



# Prendre en compte le profil génétique des patients dans la pharmacocinétique : quelles méthodes en analyse de population ?

Adrien Tessier

## ► To cite this version:

Adrien Tessier. Prendre en compte le profil génétique des patients dans la pharmacocinétique : quelles méthodes en analyse de population ?. Sciences pharmaceutiques. Université Denis Diderot Paris 7, 2016. Français. NNT : . tel-01312318

**HAL Id: tel-01312318**

**<https://www.hal.inserm.fr/tel-01312318>**

Submitted on 5 May 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**ÉCOLE DOCTORALE 393**

**ÉCOLE DOCTORALE PIERRE LOUIS DE SANTÉ PUBLIQUE À PARIS  
ÉPIDÉMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMÉDICALE**

UMR 1137 INSERM, Université Paris Diderot et Paris Nord Sorbonne Paris Cité  
Infection, Antimicrobien, Modélisation, Évolution (IAME)  
Équipe Biostatistic Modelling, Clinical Investigation and Pharmacometrics  
in Infectious Diseases

**DOCTORAT**

Spécialité : Biostatistiques

**Adrien Tessier**

**PRENDRE EN COMPTE LE PROFIL GÉNÉTIQUE DES PATIENTS DANS LA  
PHARMACOCINÉTIQUE :  
QUELLES MÉTHODES EN ANALYSE DE POPULATION ?**

**Thèse dirigée par Emmanuelle Comets**

Soutenue le 27 janvier 2016

**JURY**

Monsieur Rodolphe Thiébaut	Rapporteur
Monsieur Michel Tod	Rapporteur
Madame Emmanuelle Comets	Directrice de thèse
Madame Marylore Chenel	Co-encadrante
Madame Florence Demenais	Examinatrice
Monsieur Laurent Becquemont	Examineur



## REMERCIEMENTS

« La gratitude est non seulement la plus grande des vertus, mais c'est également la mère de tous les autres. » (Emil Cioran, XX<sup>ème</sup> siècle)

Bien qu'étant un travail personnel, une thèse ne peut s'achever que grâce au soutien d'un nombre significatif de personnes. Je tiens ici à leur exprimer mes remerciements.

Tout d'abord aux membres du jury. Je remercie Rodolphe Thiébaut et Michel Tod qui m'ont fait l'honneur d'évaluer ce travail et dont les remarques constructives ont participé à l'amélioration de ce manuscrit. Je remercie également Florence Demenais et Laurent Becquemont d'avoir accepté de participer à ce jury de thèse.

À Emmanuelle Comets, ma directrice de thèse. Je lui suis très reconnaissant de m'avoir fait profiter de ses qualités scientifiques et je la remercie également pour ses qualités humaines, la distance géographique n'ayant jamais été un obstacle à l'attention constante qu'elle m'a accordée. Je la remercie d'avoir contribué à ce que ce travail de thèse ait été aussi motivant, enrichissant et agréable.

À Marylore Chenel, ma co-encadrante de thèse. Je la remercie de m'avoir, après deux stages, renouvelé sa confiance en me permettant de faire cette thèse dans des conditions idéales, grâce non seulement au financement de la thèse, mais aussi à son soutien scientifique et personnel sans faille. Je la remercie enfin de m'avoir accueilli dans son équipe pour la dernière année de thèse.

À Julie Bertrand, qui complète cet encadrement. Je la remercie pour son soutien précieux et constant, et ceci malgré la distance. Son expertise a été un apport majeur sans laquelle cette thèse n'aurait pas été la même. Je la remercie d'avoir su m'aider pour la bonne mise en œuvre de ces travaux.

À France Mentré, directrice de l'équipe BIPID au sein de l'unité IAME UMR 1137, tout d'abord pour m'avoir accueilli dans son équipe. Je la remercie de fournir à l'ensemble des doctorants des conditions de travail optimales. Je la remercie également de nous inculquer une rigueur scientifique et d'inscrire une dynamique motivante au sein de cette équipe en nous poussant notamment à participer à des congrès et à mettre en avant nos travaux. Pour



finir, je la remercie, au-delà de son statut de directrice, pour tout ce qu'elle a pu faire sur le plan humain, autant avant que pendant la thèse. Mon passage dans cette équipe restera une pierre angulaire de ma carrière professionnelle et scientifique.

À l'ensemble de l'équipe BIPID. D'abord à Houda pour tout ce qu'elle met en œuvre pour le bien être de cette équipe. À Marie pour son aide et sa bonne humeur tout au long de la thèse. À tous mes collègues doctorants, anciens ou actuels : Blaise, Camille, Cédric, Charles, Cyrielle, Giulia, Jean, Josselin, Julien, Minh, Philippine, Sarah, Simon, Solène, Steven, Thu, Thu-Thuy, Tram et Vincent. Aux autres membres de l'équipe : François, Hervé, Jérémie, Jimmy, Marie-Karelle, Minerva, Pauline, Sebastian. Merci pour votre accueil et votre sympathie. Vous avez participé à créer cette ambiance si agréable. Pour finir à l'équipe DeSCID avec qui on partage bons moments, en plus de partager un étage.

À l'ensemble de l'équipe de Pharmacocinétique Clinique et de Pharmacométrie, au sein d'IRIS. À Anne, Charlotte, Emilie, Isabelle, Julien, Kathryn, Maud, Sogo, Stéphan, Sylvain et Tanguy. Aux personnes rencontrées pendant ces trois années : Alexia, Anaïs, Camille, Caroline, Clémence, Constance, Jacques, Katie, Lucie, Manon, Marc, Marylène, Pauline et Quentin. Merci pour votre soutien, votre sympathie, et pour votre accueil chaleureux. Une attention spéciale à Karl qui m'a encadré pendant l'année que j'ai passé dans l'équipe, et à Sophie avec qui nous nous sommes soutenus mutuellement pour la dernière ligne droite de nos thèses respectives. Pour finir à Bernard Walther et Laurent Ripoll pour leur aide sur mes différents projets.

À Marc Lavielle pour son aide sur le troisième projet.

À mes anciens collègues représentants des doctorants : Enora, Jacques-Emmanuel, Laure-Amélie et Thibaut.

À ma famille. À mes parents, qui m'ont supporté et m'ont encouragé sans cesse pendant toutes ces années d'études (*sans faute cette fois...*). À mes frères, qui m'ont également supporté, d'une manière différente...

À mes amis. Certains sont déjà cités plus haut, je me permettrais donc de ne pas les citer à nouveau. À Abdou, Audrey, Amaury, Anne-Laure, Aziz, Bénédicte, Camille, Cécile, Corentine, Elodie, Estelle, Florent, Floris, JV, Justine, Laurie, Lévi, Mailys, Marion, Olivia, Ophélie, PH, Sam, Romain, Thierry. Aux membres de Pharmazik.

À Etty

« Racontez-moi, et j'oublierai,  
Démontrez-moi, et je me souviendrai,  
Impliquez-moi, et je comprendrai. »  
(Xun Kuang, III<sup>ème</sup> siècle avant J.C.)

# TABLE DES MATIÈRES

REMERCIEMENTS .....	i
TABLE DES MATIÈRES .....	v
PRODUCTIONS SCIENTIFIQUES LIÉES À LA THÈSE .....	vii
GLOSSAIRE .....	ix
INTRODUCTION .....	1
1.1. La Pharmacocinétique .....	2
1.1.1. Définition .....	2
1.1.2. Le processus ADME .....	3
1.1.3. Les enzymes du métabolisme, les transporteurs, et leur implication dans la variabilité des processus pharmacocinétiques .....	4
1.1.4. Les méthodes d'analyse de profils pharmacocinétiques individuels .....	9
1.1.5. Les modèles non linéaires à effets mixtes .....	16
1.1.6. Rôle de la modélisation en pratique clinique et dans le développement du médicament .....	26
1.2. La Pharmacogénétique .....	27
1.2.1. Définitions .....	27
1.2.2. Les polymorphismes génétiques .....	28
1.2.3. Les analyses pharmacogénétiques .....	39
1.2.4. La pharmacogénétique en pratique clinique .....	49
1.2.5. La pharmacogénétique dans le développement du médicament .....	53
PRÉSENTATION DU CAS RÉEL .....	59
2.1. Données .....	59
2.2. Analyse pharmacocinétique .....	62
OBJECTIFS DE LA THÈSE .....	67

COMPARAISON DES MÉTHODES D'ESTIMATION DES PHÉNOTYPES PHARMACOCINÉTIQUES ET DE TESTS D'ASSOCIATION POUR LES ANALYSES PHARMACOGÉNÉTIQUES .....	69
4.1. Résumé .....	69
4.2. Article 1 (publié) .....	71
4.3. Matériel supplémentaire .....	84
IMPORTANCE DU PROTOCOLE DES ÉTUDES PHARMACOCINÉTIQUES DANS LA PUISSANCE DES ANALYSES PHARMACOGÉNÉTIQUES : INTÉRÊTS DES PROCOTOLES COMBINÉS .....	98
5.1. Résumé .....	98
5.2. Article 2 (accepté).....	100
5.3. Matériel supplémentaire .....	110
PRISE EN COMPTE DE L'INTÉGRALITÉ DE LA DISTRIBUTION CONDITIONNELLE DANS LES TESTS D'ASSOCIATION .....	131
6.1. Introduction .....	131
6.2. Matériels et méthodes .....	132
6.2.1. Étude de simulation .....	132
6.2.2. Obtention des tirages dans la distribution conditionnelle .....	134
6.2.3. Évaluation .....	135
6.3. Résultats .....	135
6.3.1. Erreur de type I.....	135
6.3.2. <i>Shrinkage</i> .....	136
6.3.3. Probabilité de detection.....	138
6.4. Discussion .....	139
DISCUSSION GÉNÉRALE .....	142
7.1. Discussion des résultats.....	142
7.2. Recommandations pour les analyses pharmacogénétiques.....	148
BIBLIOGRAPHIE.....	151

# PRODUCTIONS SCIENTIFIQUES LIÉES À LA THÈSE

## Article publié

Tessier, A., Bertrand, J., Chenel, M., Comets, E., 2015. Comparison of nonlinear mixed effects models and noncompartmental approaches in detecting pharmacogenetic covariates. AAPS J. 17, 597–608.

## Article accepté

Tessier, A., Bertrand, J., Chenel, M., Comets, E., 2015. Combined analysis of phase I and phase II data to enhance the power of pharmacogenetic tests. CPT Pharmacometrics Syst Pharmacol. Accepté le 11 décembre 2015

## Communications orales

Tessier, A., Bertrand, J., Fouliard, S., Chenel, M., Comets, E., Septembre 2013. Inclusion of genetic information in population pharmacokinetic models. ESPT 2013, second conference: Pharmacogenomics, Lisbonne, Portugal.

Tessier, A., Bertrand, J., Chenel, M., Comets, E., Juin 2015. Modelling pharmacogenetic data in population studies during drug development. 24<sup>th</sup> Population Group Approach in Europe conference (PAGE), Hersonissos, Crète, Grèce.

Tessier, A., Bertrand, J., Chenel, M., Comets, E., Octobre 2015. How to integrate pharmacogenetics in population PK modeling. The GMP (Groupe de Métabolisme et Pharmacocinétique) Open Meeting, Paris, France.

## Communications affichées

Tessier, A., Bertrand, J., Fouliard, S., Comets, E., Chenel, M., Avril 2013. Methods for the inclusion of genetic information in population pharmacokinetic models. 8<sup>ème</sup> Congrès de Physiologie, Pharmacologie et Thérapeutique, Angers, France.

Tessier, A., Bertrand, J., Fouliard, S., Comets, E., Chenel, M., Juin 2013. High-throughput genetic screening and pharmacokinetic population modeling in drug development. 22<sup>th</sup> Population Group Approach in Europe conference (PAGE), Glasgow, Royaume-Uni.

Tessier, A., Bertrand, J., Chenel, M., Comets, E., Octobre 2013. Contribution of nonlinear mixed effects models and penalised regression approaches in pharmacogenetic population association studies. The GMP (Groupe de Métabolisme et Pharmacocinétique) Open Meeting, Paris, France.

Tessier, A., Bertrand, J., Chenel, M., Comets, E., Juin 2014. Contribution of nonlinear mixed effects models and penalised regression approaches in pharmacogenetic population studies. 23<sup>th</sup> Population Group Approach in Europe conference (PAGE), Alicante, Espagne.

Tessier, A., Bertrand, J., Chenel, M., Comets, E., Octobre 2014. Mixed designs in pharmacogenetics: a simulation study for clinical drug development. DMDG/GMP Joint Meeting, Paris, France.

Brendel, K., Tessier, A., Chenel, M., Juin 2015. How to consider microdosing data in a population pharmacokinetic analysis? 24<sup>th</sup> Population Group Approach in Europe conference (PAGE), Hersonissos, Crète, Grèce.

## GLOSSAIRE

$\alpha$	Erreur de type I par test
$\beta$	Coefficient d'effet
A	Adénine
ADME	Absorption, distribution, métabolisme, excrétion
ADN	Acide désoxyribonucléique
AMM	Autorisation de mise sur le marché
ARN	Acide ribonucléique
AUC	Aire sous la courbe ( <i>Area under the curve</i> )
C	Cytosine
CL	Clairance d'élimination
CYP	Cytochrome P450
CWRES	Résidus conditionnels de population pondérés ( <i>Conditional weighted residuals</i> )
DV	Variable dépendante ( <i>Dependent variable</i> )
<i>e.g.</i>	Par exemple ( <i>exempli gratia</i> )
EBE	Estimations bayésiennes des paramètres individuels ( <i>Empirical Bayesian Estimates</i> )
EMA	Agence européenne du médicament ( <i>European Medicines Agency</i> )
F	Biodisponibilité
FDA	Agence américaine du médicament ( <i>Food and Drug Administration</i> )
FRAC	Fraction de la dose
FWER	Erreur de type I globale ( <i>Family wise error rate</i> )
G	Guanine
GWAS	Étude génome entier ( <i>Genome-wide association study</i> )
<i>i.e.</i>	c'est-à-dire ( <i>id est</i> )
IIV	Variabilité interindividuelle ( <i>Interindividual variability</i> )
IOV	Variabilité interoccasion ( <i>Interoccasion variability</i> )
IPRED	Prédictions individuelles ( <i>Individual predictions</i> )
IRIS	Institut de recherches internationales Servier
IWRES	Résidus individuels pondérés ( <i>Individual weighted residuals</i> )
$k_a$	Constante de vitesse d'absorption d'ordre 1
$k$	Constante de vitesse d'élimination
Lasso	<i>Least absolute shrinkage and selection operator</i>
LD	Déséquilibre de liaison ( <i>Linkage disequilibrium</i> )
MAF	Fréquence de l'allèle mineur ( <i>Minor allele fraction</i> )
MCMC	Méthode de Monte Carlo par chaînes de Markov
MNLEM	Modèle non linéaire à effets mixtes
NCA	Analyse non compartimentale ( <i>Noncompartmental analysis</i> )
NPDE	Erreurs de prédiction de distribution normalisées ( <i>Normalised prediction distribution errors</i> )



PD	Pharmacodynamie ( <i>Pharmacodynamics</i> )
PGt	Pharmacogénétique ( <i>Pharmacogenetics</i> )
PGx	Pharmacogénomique ( <i>Pharmacogenomics</i> )
PK	Pharmacocinétique ( <i>Pharmacokinetics</i> )
PRED	Prédictions de population ( <i>Population predictions</i> )
Q	Clairance intercompartimentale
RSE	Erreur standard relative ( <i>Relative standard error</i> )
SE	Erreur standard ( <i>Standard error</i> )
<i>Sh</i>	Métrique quantifiant la régression vers la moyenne ( <i>Shrinkage</i> )
SNP	Substitution nucléotidique ( <i>Single nucleotide polymorphism</i> )
T	Thymine
$t_{1/2}$	Temps de demi-vie
Tk0	Durée de l'absorption d'ordre 0
Tlag	Temps de latence ( <i>Lag time</i> )
V	Volume du compartiment
Vd	Volume de distribution
VPC	<i>Visual predictive check</i>

# Chapitre 1

## INTRODUCTION

La personnalisation de la médecine est un enjeu majeur dans la prise en charge des maladies, afin de proposer le bon traitement avec le bon schéma posologique, au bon groupe de patients. En effet, la réponse aux médicaments varie d'un patient à l'autre (Wilkinson, 2005), ou d'une administration à l'autre chez un même patient. Et ce malgré un diagnostic, *i.e.* une maladie et un niveau de sévérité, identique (Burton et al., 2005). Par réponse, on entend aussi bien l'effet thérapeutique que les possibles effets toxiques d'un médicament.

La variabilité dans la réponse à un traitement peut être décomposée en une variabilité pharmacocinétique, *i.e.* le devenir du médicament dans l'organisme, et une variabilité pharmacodynamique, *i.e.* l'effet du médicament sur l'organisme. De nombreux facteurs peuvent être à l'origine de cette variabilité (Houin, 1998; Tozer and Rowland, 2006); parmi eux on retrouve des facteurs physiopathologiques (âge, poids, insuffisance rénale ou hépatique...), pharmacologiques (co-administration de médicaments...), environnementaux (régime alimentaire, exposition à des infections modifiant le système immunitaire...) mais également génétiques (Allorge and Lorient, 2004).

Nous nous intéressons dans cette thèse aux facteurs génétiques pouvant expliquer une partie de la variabilité de la réponse au traitement entre les patients et plus précisément à l'effet de variants génétiques sur la pharmacocinétique. En collaboration avec l'industrie pharmaceutique (IRIS, Institut de Recherches Internationales Servier), nous nous sommes intéressés plus particulièrement à la méthodologie dans le développement clinique de nouveaux médicaments permettant de mettre en évidence l'implication de facteurs génétiques dans la variabilité pharmacocinétique de ces nouvelles molécules.

Dans cette introduction nous définirons le concept de pharmacocinétique. Nous développerons notamment les méthodes d'estimation des paramètres pharmacocinétiques et de leur variabilité, en s'intéressant plus particulièrement aux modèles non linéaires à effets mixtes. Puis nous définirons le concept de pharmacogénétique. Nous développerons

les études et les méthodes d'analyse d'associations pharmacogénétiques, avant d'évoquer la place de la pharmacogénétique dans la pratique clinique et dans le développement du médicament.

## **1.1. La Pharmacocinétique**

### **1.1.1. Définition**

La pharmacocinétique (PK), terme introduit en 1953 (Dost, 1953), étudie « ce que fait le corps au médicament », dans une formule vulgarisée (Wagner, 1981). La définition proposée par Gibaldi et Levy est la suivante : « La pharmacocinétique étudie et caractérise l'évolution au cours du temps de l'absorption, la distribution, le métabolisme et l'excrétion des médicaments, ainsi que la relation entre ces processus et l'intensité et l'évolution au cours du temps des effets thérapeutiques et indésirables de ces mêmes médicaments. La pharmacocinétique implique l'utilisation de méthodes mathématiques dans un contexte physiologique et pharmacologique » (traduction adaptée de (Gibaldi and Levy, 1976)).

On trouve dans cette définition la notion de processus dynamiques, *i.e.* évoluant au cours du temps. La pharmacocinétique consiste à étudier qualitativement et quantitativement le devenir des médicaments, quantifié par l'évolution des concentrations du médicament dans le temps et dans différents milieux (*i.e.* le sang, les urines...). On s'intéresse principalement aux concentrations plasmatiques représentant la quantité de médicament dans la circulation générale, mais on peut également étudier les concentrations au niveau périphérique (par exemple dans le liquide céphalorachidien afin d'étudier la part de médicament atteignant le cerveau) ou la quantité de médicament éliminée dans les urines ou les fèces.

On trouve également dans la définition de Gibaldi et Levy la relation concentrations - effet démontrée pour la première fois en 1932 par Widmark qui mis en évidence la relation entre la concentration sanguine de l'alcool et les effets psychoactifs de ce dernier (Widmark, 1932). La pharmacodynamie (PD) étudie les effets thérapeutiques et toxiques d'un médicament sur l'organisme. Ces effets varient en fonction de la marge thérapeutique et des concentrations du médicament présentes dans l'organisme. On parle donc de relation pharmacocinétique/pharmacodynamique (PK/PD).

L'utilisation de méthodes mathématiques pour étudier la pharmacocinétique se traduit dès 1937 par un premier modèle compartimental, proposé par Teorell, où l'organisme était

représenté par cinq compartiments (Teorell, 1937). Ces approches seront développées dans le paragraphe 1.1.4.2.

Le parcours du médicament dans l'organisme est classiquement décomposé en quatre grandes phases nommées alors résorption, distribution, consommation et élimination (Teorell, 1937). Ces termes ont par la suite été modifiés pour devenir l'acronyme ADME que l'on connaît aujourd'hui (Nelson, 1961), signifiant absorption, distribution, métabolisme et excrétion (les deux dernières phases peuvent être groupées le terme générique d'élimination, on parle alors de processus ADE) (Gabrielson and Weiner, 2007; Rowland and Tozer, 2011).

### **1.1.2. Le processus ADME**

La phase d'absorption représente le passage du site d'administration à la circulation générale et dépend donc de la voie d'administration du médicament. Dans le cas d'une administration intraveineuse (i.v.), le médicament est administré directement dans la circulation générale, il n'y a donc pas d'absorption. La majorité des médicaments est administrée par voie orale et, dans ce cas, l'absorption représente le passage de la lumière intestinale à la circulation générale, ce qui nécessite notamment le passage des molécules à travers la barrière intestinale. La dose de médicament administrée par voie orale n'est généralement pas retrouvée en totalité dans la circulation générale (à l'inverse d'une administration i.v.), en raison de plusieurs facteurs. En effet, une partie de cette dose n'est jamais absorbée car elle ne s'est pas totalement libérée de sa forme galénique (*e.g.* comprimé) en raison d'une mauvaise solubilité, et/ou parce qu'elle ne s'est pas totalement dissoute dans le tube digestif. Le passage du tube digestif à la circulation générale peut également ne pas se faire en totalité à cause de la faible perméabilité de la molécule, mais également en raison de phénomènes de saturation de l'absorption ou d'efflux (dont nous parlerons dans le paragraphe 1.1.3). Cette fraction de la dose non absorbée est directement éliminée dans les fèces. Une partie de la dose peut également être métabolisée au niveau des cellules de la membrane intestinale (entérocytes) puis au niveau du foie, les molécules absorbées passant d'abord par la veine porte reliant la membrane intestinale au foie avant d'arriver à la circulation générale. C'est l'effet de premier passage (intestinal puis hépatique). On quantifie la fraction de médicament réellement absorbée en calculant sa biodisponibilité symbolisée par le paramètre F (exprimé en pourcentage).

Une fois dans la circulation générale, le médicament peut se lier de façon dynamique aux protéines plasmatiques (albumine, alpha-glycoprotéines...). La fraction libre se distribue dans l'ensemble de l'organisme (organes et tissus) en fonction de différents facteurs (perméabilité des tissus, débit sanguin de perfusion des tissus, propriétés physico-chimiques de la molécule). On définit le paramètre Vd comme le volume de distribution reliant la quantité totale de médicament présente dans l'organisme et les concentrations observées au niveau plasmatique. Ainsi pour une même quantité de médicament, une molécule se distribuant davantage dans les tissus est caractérisée par des concentrations plasmatiques plus faibles et donc un volume de distribution plus élevé. De ce fait les volumes de distribution mesurés n'ont pas de réalité biologique et peuvent prendre des valeurs bien supérieures au volume physiologique (e.g. l'amiodarone a un volume de distribution chez l'Homme de 62 L/kg (Holt et al., 1983), pour un volume sanguin physiologique de 4 à 6 L).

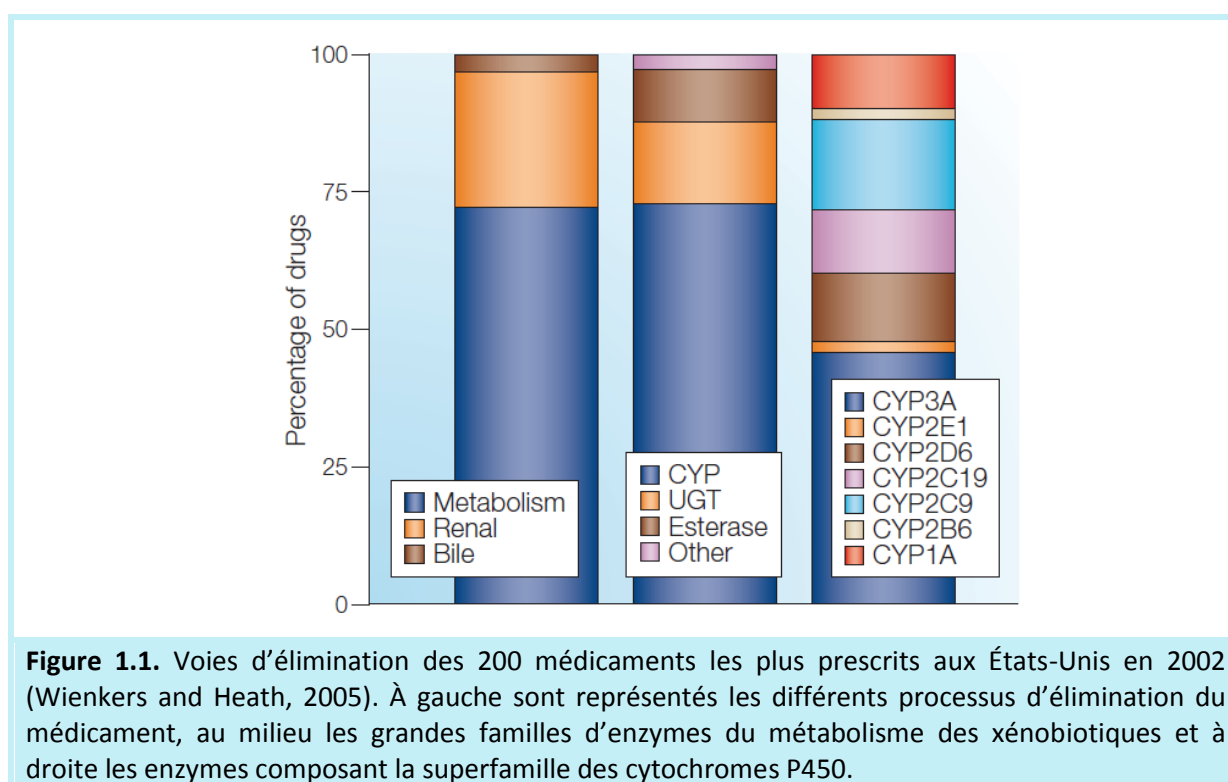
En se distribuant, le médicament atteint les organes responsables de son élimination. On divise l'élimination en deux processus, le métabolisme du médicament (molécule mère) et l'excrétion de la molécule mère elle-même ou de ses métabolites. La métabolisation est la transformation chimique de la molécule mère en un ou plusieurs métabolites. C'est la voie principale d'élimination des médicaments (**Figure 1.1**). Elle se fait essentiellement au niveau du foie qui comprend une quantité très importante d'enzymes. L'intérêt du métabolisme est d'obtenir des métabolites plus solubles que le médicament lui-même, qui pourront être excrétés plus facilement. L'excrétion est l'élimination du médicament ou de ses métabolites dans les fluides biologiques, généralement dans l'urine à travers les reins ou la bile à travers le foie. Par ces deux processus l'organisme se protège face à des composés étrangers, les xénobiotiques, dont font partie les médicaments. Pour quantifier l'élimination du médicament on définit sa clairance (CL), qui correspond au volume sanguin totalement épuré du médicament par unité de temps, soit un débit exprimé en  $\text{volume} \cdot \text{temps}^{-1}$ .

### **1.1.3. Les enzymes du métabolisme, les transporteurs, et leur implication dans la variabilité des processus pharmacocinétiques**

De très nombreuses protéines interviennent dans la pharmacocinétique du médicament et ce à toutes les phases du processus ADME. On compte parmi elles les enzymes du métabolisme, les transporteurs et les récepteurs nucléaires.

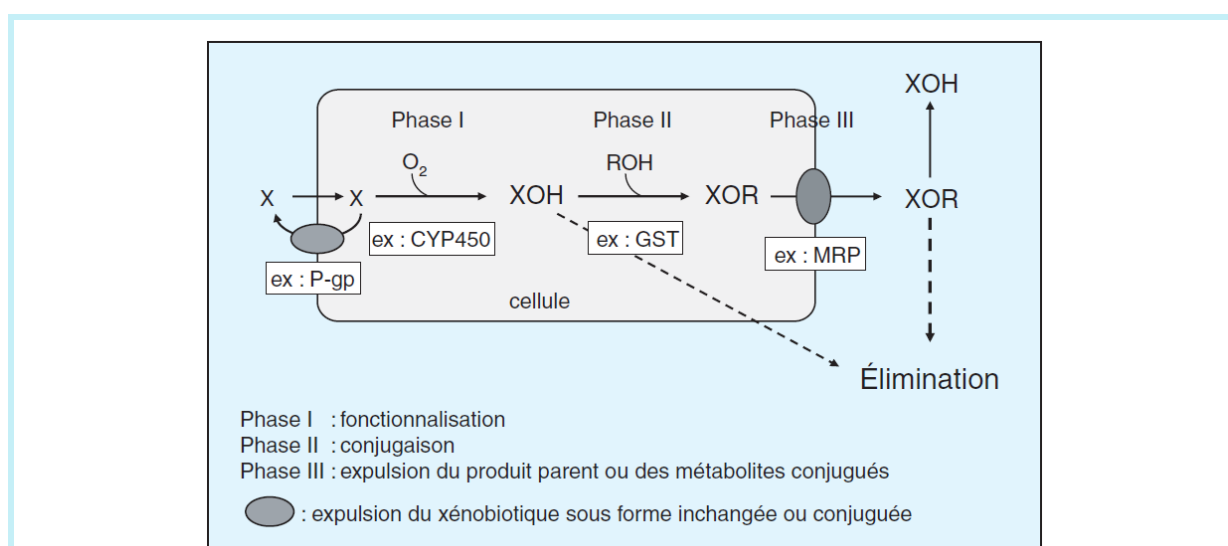
Les enzymes du métabolisme ont été très largement étudiées et font l'objet de nombreuses publications. On distingue les enzymes catalysant des réactions de phase I (oxydation,

réduction ou hydrolyse) et les enzymes catalysant des réactions de phase II (conjugaison). Bien qu'il existe différents types d'enzymes de phase I (e.g. des estérases responsables de réactions d'hydrolyse ou des réductases catalysant des réductions), ce sont les réactions d'oxydation qui ont été le plus étudiées, catalysées principalement par les enzymes de la superfamille des cytochromes P450 (CYP) à l'origine de la métabolisation de la majorité des médicaments (**Figure 1.1**). On compte 57 gènes chez l'Homme codant pour des CYP classifiés en familles, sous-familles et isoformes. Les isoformes les plus impliquées dans le métabolisme des médicaments sont les CYP3A4, CYP2D6, CYP2C9, CYP2C19 et CYP1A2 (Galetin et al., 2010). Les enzymes de phase II, des transférases, vont être responsables de la conjugaison de nouveaux groupements chimiques sur les molécules. De nombreuses familles catalysent ces réactions et sont elles-mêmes composées de plusieurs isoformes. Les familles les plus importantes sont les UDP-glucuronotransférases (UGT) comptant 17 isoformes chez l'Homme, les sulfo-transférases (SLT) ou les N-acétyltransférases (NAT).



Les transporteurs sont des protéines membranaires assurant le passage du médicament à travers la membrane cellulaire (Giacomini and Huang, 2013). L'implication de ces transporteurs peut être divisée en deux actions (**Figure 1.2**) selon qu'ils assurent l'entrée de la molécule dans la cellule (influx), ou sa sortie de la cellule (efflux). Deux grandes familles

composent les transporteurs : la famille des transporteurs *Solute Carrier* (SLC) et la famille des transporteurs *ATP-Binding Cassette* (ABC). Il existe huit familles de transporteurs SLC impliquées dans le transport des médicaments dont la famille des OATP et OAT transportant des anions, et des OCT transportant des cations. Trois familles de transporteurs ABC sont responsables du transport des médicaments : la famille ABCB, ABCC et ABCG. On trouve notamment dans la famille ABCB, la glycoprotéine P (P-gp ou ABCB1) qui joue un rôle majeur en pharmacocinétique (Marzolini et al., 2004). En effet, ce transporteur d'efflux peut au niveau des entérocytes empêcher le passage de la barrière intestinale, diminuant ainsi l'absorption du médicament. Au niveau de la barrière hémato-encéphalique, il empêche l'entrée des xénobiotiques dans le cerveau, le protégeant de substances potentiellement toxiques. Il facilite également le passage du sang vers la bile au niveau des hépatocytes et vers les urines au niveau du rein, augmentant ainsi l'excrétion des médicaments.



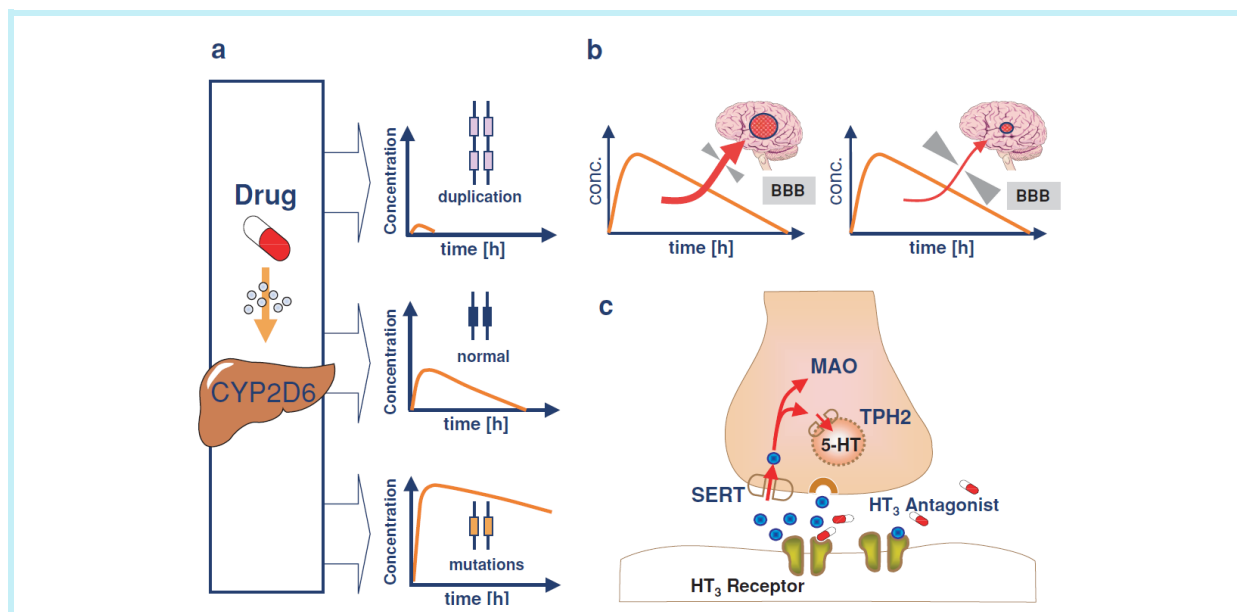
**Figure 1.2.** Métabolisme des médicaments (Allorge and Lorient, 2004). De gauche à droite, le médicament  $X$  entre dans la cellule. Cette entrée peut être empêchée par un transporteur d'efflux, ici la P-gp (phase 0). Une fois dans la cellule,  $X$  subit une réaction de phase I catalysée par un CYP puis une réaction de phase II catalysée ici par une glutathion S-transférase (GST). La sortie du métabolite  $XOR$  se fait à l'aide d'un transporteur (phase III) puis le métabolite est excrété.

Les récepteurs nucléaires comme PXR (*pregnane X receptor*) ou CAR (*constitutive androstane receptor*) sont des régulateurs transcriptionnels actifs dans le noyau des cellules. Ils interviennent au niveau génétique dans la synthèse des autres protéines, dont les enzymes du métabolisme et les transporteurs (Lamba et al., 2008), en modulant l'activité de gènes cibles. La variabilité dans l'activité de ces récepteurs nucléaires peut donc modifier aussi

bien le métabolisme que le transport du médicament. Ceci rend difficilement interprétable l'effet d'une variation de ces protéines sur la pharmacocinétique.

L'activité des protéines citées ci-dessus varie en fonction de plusieurs facteurs. Par exemple dans la population pédiatrique, l'activité de certains cytochromes varie en fonction de l'âge (de Zwart et al., 2004). Ces différences sont surtout observées à la naissance et pendant les premiers mois de la vie où plusieurs CYP sont immatures (CYP3A4, CYP2D6, CYP1A2...), conduisant ainsi à des capacités métaboliques réduites par rapport à l'adulte (Johnson et al., 2006). Certaines pathologies peuvent aussi modifier l'activité des protéines. C'est notamment le cas de pathologies touchant le foie ou le rein, organes responsables de l'élimination d'une grande partie des médicaments. Certains médicaments inhibent ou induisent l'activité des enzymes du métabolisme ou des transporteurs et sont à l'origine d'interactions médicamenteuses. Enfin, les variations génétiques peuvent être responsables d'une diminution ou d'une augmentation de l'activité enzymatique, se traduisant par une variabilité dans l'exposition des individus aux médicaments. Pour exemple, la **Figure 1.3** illustre l'influence des variations génétiques sur l'exposition au niveau central et périphérique au tropisétron, un antagoniste du récepteur à la sérotonine HT<sub>3</sub> utilisé comme antiémétique (Eichelbaum et al., 2006). Ce médicament est métabolisé par le CYP2D6. Pour une même dose, les patients avec une activité enzymatique du CYP2D6 augmentée par duplication du gène ont des expositions systémiques bien inférieures aux patients porteurs d'un gène normal (nous employons ici un vocabulaire simple avant de définir les termes spécifiques à la génétique dans le paragraphe 1.2). À l'inverse les porteurs d'une mutation responsable d'une diminution de l'activité enzymatique ont des expositions plus élevées (**Figure 1.3a**). Le tropisétron est également substrat de la P-gp. Au niveau de la barrière hémato-encéphalique, les patients porteurs de mutations responsables d'une augmentation de l'activité d'efflux de la P-gp ont des concentrations dans le système nerveux central (le site d'action) inférieures (**Figure 1.3b**), et ce pour des expositions systémiques similaires aux porteurs du gène normal.





**Figure 1.3.** Illustration de l'impact des variations génétiques sur la pharmacocinétique des médicaments (Eichelbaum et al., 2006).

- Exposition systémique de l'organisme après administration d'une dose de tropisétro, en fonction des variations génétiques du CYP2D6.
- Passage du tropisétro au niveau périphérique (cerveau) à travers la barrière hémato-encéphalique, en fonction de l'activité de la P-gp liée aux variations génétiques.
- Effet du tropisétro en fonction de variations génétiques modifiant l'activité des récepteurs et les concentrations des neurotransmetteurs (pharmacodynamie).

À ce jour 30 à 50% des médicaments sur le marché sont métabolisés par des enzymes polymorphiques, *i.e.* présentant des variations génétiques (Eichelbaum et al., 2006; Evans and Relling, 1999), et 80% des informations sur la variabilité d'origine génétique contenues dans la notice des médicaments concernent ces enzymes du métabolisme (Frueh et al., 2008).

Quantitativement, l'impact des polymorphismes des transporteurs est moins connu, comparé aux enzymes du métabolisme qui ont été très largement étudiées. C'est notamment dû au fait que les polymorphismes des transporteurs peuvent modifier non pas la cinétique au niveau systémique, mais plutôt au niveau périphérique (*e.g.* au niveau du site d'action), ce qui est plus complexe à mesurer. Malgré cela, des exemples d'associations entre des polymorphismes de transporteurs et la pharmacocinétique de médicaments ont été mis en évidence (Comets et al., 2007; de Keyser et al., 2014).

La pharmacogénétique, que nous développerons en détail ultérieurement, étudie l'association entre la variabilité pharmacocinétique et les variations génétiques. Dans ce cas, la variable dépendante, *i.e.* le phénotype, peut être l'ensemble des paramètres PK d'une

population. Il est donc nécessaire de correctement caractériser la variabilité des processus PK lors de la détermination de ces paramètres. Pour cela différentes approches sont disponibles.

#### **1.1.4. Les méthodes d'analyse de profils pharmacocinétiques individuels**

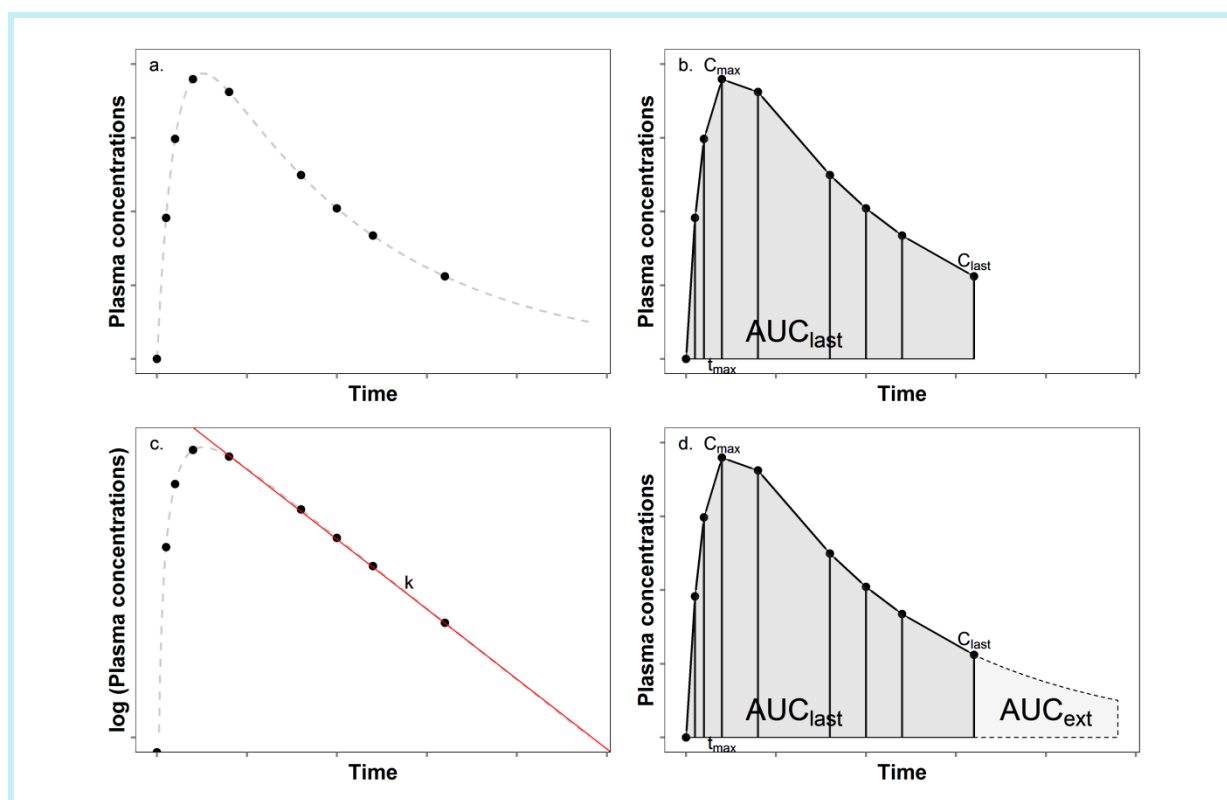
La pharmacométrie regroupe l'ensemble des méthodes d'analyse ayant pour objet l'évaluation quantitative et qualitative de la pharmacocinétique et de la pharmacodynamie d'un médicament (définition adaptée du dictionnaire de l'Académie nationale de Pharmacie). Dans ce paragraphe, nous nous intéressons à l'étude de la pharmacocinétique au niveau individuel. Dans le paragraphe suivant, nous montrerons comment l'étude au niveau de la population permet de mieux quantifier la variabilité des processus PK.

Il existe deux grandes approches pour étudier le profil pharmacocinétique d'un individu obtenu à partir des concentrations mesurées : l'analyse non compartimentale (NCA pour *noncompartmental analysis*) ou l'analyse compartimentale, *i.e.* la modélisation.

##### **1.1.4.1. L'analyse non compartimentale**

La NCA est historiquement la première approche utilisée pour étudier la pharmacocinétique d'un médicament. La pharmacocinétique d'une molécule est résumée par des paramètres secondaires estimés graphiquement à partir des concentrations mesurées au cours du temps (Gabrielson and Weiner, 2007). C'est une approche uniquement descriptive utilisant des méthodes de calculs simples. La **figure 1.4** illustre le calcul de différents paramètres PK à partir des concentrations mesurées chez un individu (**Figure 1.4a**). L'aire sous la courbe (AUC pour *area under the curve*) quantifie l'exposition totale de l'organisme au médicament. Elle est généralement estimée par NCA en plusieurs étapes. Dans un premier temps, l'aire jusqu'à la dernière concentration observée ( $AUC_{last}$ ) est estimée par la méthode des trapèzes (**Figure 1.4b**). Cette observation correspond à la dernière concentration mesurée selon le protocole ( $C_{last}$ ), ou à la dernière concentration mesurable par la méthode analytique (*i.e.* au-dessus de la limite de quantification). Ensuite, l'élimination du médicament est estimée par la pente terminale ( $k$ ) à partir des dernières concentrations observées (**Figure 1.4c**). La pharmacocinétique est généralement caractérisée par des processus d'élimination d'ordre 1, la décroissance des concentrations est donc exponentielle. En représentant un profil PK sur une échelle semi-logarithmique, on observe une décroissance linéaire des concentrations au cours du temps, dont on peut estimer la pente. Lorsque plusieurs pentes sont observées sur un profil, la pharmacocinétique de la molécule est composée de différentes phases

mélangeant des processus de distribution et d'élimination. La dernière décroissance de pente  $k$  représente uniquement la phase d'élimination du médicament. L'AUC non observée après la dernière concentration mesurée est extrapolée par le ratio de la concentration  $C_{last}$  sur la pente terminale  $k$  ( $AUC_{ext}$ ). L'AUC totale est obtenue par la somme de l' $AUC_{last}$  et de l' $AUC_{ext}$  (Figure 1.4d).



**Figure 1.4.** Illustration du calcul des paramètres pharmacocinétiques en analyse non compartimentale.

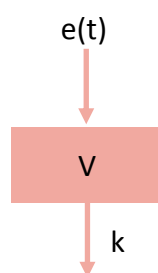
- Les concentrations plasmatiques mesurées chez un individu en fonction du temps sont représentées par les points. Les concentrations non observées sont représentées par la courbe pointillée, décrivant le profil PK complet de l'individu.
- Détermination de la concentration maximale ( $C_{max}$ ) et du temps à cette concentration ( $t_{max}$ ), de la dernière concentration mesurée ( $C_{last}$ ) et de l'aire sous la courbe jusqu'à la dernière observation ( $AUC_{last}$ ) par la méthode des trapèzes.
- Représentation des concentrations en échelle logarithmique et détermination de la pente terminale ( $k$ ) représentée par la droite rouge. La courbe pointillée représente les concentrations non observées.
- Extrapolation de l'aire sous la courbe ( $AUC_{ext}$ ) afin d'estimer l'aire de l'ensemble du profil PK.

À partir de la pente  $k$ , peut également être calculée la demi-vie du médicament ( $t_{1/2} = \frac{\ln 2}{k}$ ), paramètre qui représente le temps nécessaire pour diminuer de moitié la quantité de produit dans l'organisme. Ce paramètre permet de déterminer le temps à partir duquel le

médicament est totalement éliminé, usuellement fixé à  $5 t_{1/2}$ , et ainsi de choisir le schéma d'administration. D'autres paramètres peuvent être renseignés comme la concentration maximale observée ( $C_{\max}$ ) et le temps auquel cette concentration maximale est observée ( $t_{\max}$ ). Ces paramètres PK secondaires n'ont pas de signification physiologique mais peuvent être utilisés pour dériver des paramètres primaires, *i.e.* physiologiques, mentionnés au paragraphe 1.1.2 : i) la biodisponibilité  $F$  est calculée par le rapport des AUC entre une dose de référence administrée par voie i.v. et la même dose administrée par voie orale, ii) la clairance peut être calculée par la formule  $CL = \frac{Dose}{AUC}$  et iii) le volume de distribution peut être calculé comme étant  $Vd = \frac{Dose}{C_0}$  où  $C_0$  représente la concentration mesurée au temps 0 après une administration i.v. bolus.

### 1.1.4.2. L'analyse compartimentale

Cette approche repose sur un modèle mathématique pour décrire les données observées. Dans cette approche, l'organisme est décomposé en un système de compartiments, où un compartiment représente un espace cinétiquement homogène. Dans l'approche empirique, l'organisme est résumé en un minimum de compartiments, reliés entre eux par des constantes de transferts (**Figure 1.5**) (Rowland and Tozer, 2011). Ces constantes de transferts symbolisent les échanges réversibles ou irréversibles entre les compartiments et s'expriment en temps<sup>-1</sup>. Elles décrivent des processus d'ordre 1, proportionnels à la quantité de produit présente à un instant  $t$ . Dans certains cas, on observe une saturation des processus PK. Des cinétiques d'ordre 0 indépendantes de la quantité (uniquement pour l'absorption), ou des relations de type Michaelis-Menten (Michaelis et al., 2011; Michaelis and Menten, 1913) où la vitesse d'absorption ou d'élimination dépend de la quantité de façon non linéaire sont alors utilisées.

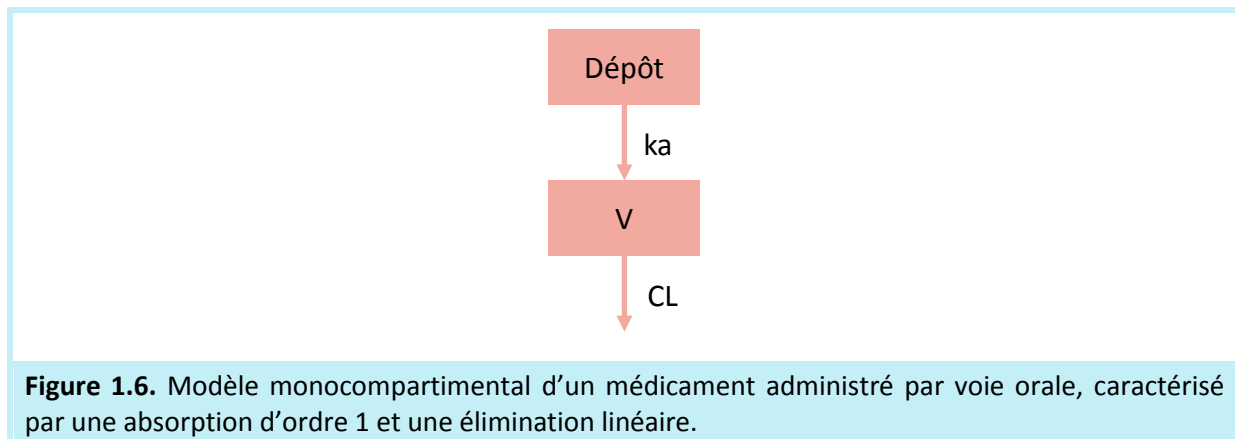


**Figure 1.5.** Modèle monocompartimental d'un médicament caractérisé par un volume  $V$ , une fonction d'entrée  $e(t)$  et une élimination linéaire de constante  $k$ .

Le modèle le plus simple ne comprend qu'un unique compartiment et suppose ainsi que la cinétique du médicament est homogène dans tout l'organisme (**Figure 1.5**). Ce compartiment est caractérisé par un volume  $V$  et la concentration de médicament dans ce compartiment est calculée par le ratio de la quantité présente à un temps  $t$  sur le volume du compartiment. La dynamique de cette quantité au cours du temps est le résultat d'un bilan de matière, *i.e.* la quantité entrant dans le compartiment à laquelle est soustraite la quantité en sortant. Ainsi, ce compartiment peut être caractérisé par une équation différentielle :

$$\frac{dA(t)}{dt} = e(t) - k \cdot A(t) \quad (1.1)$$

où  $A(t)$  est l'évolution de la quantité de médicament dans le compartiment, qui dépend d'une fonction d'entrée  $e(t)$  et de la constante d'élimination  $k$ . Cette fonction d'entrée est différente selon le type d'administration. Par exemple, pour une administration i.v. bolus, au temps 0 la totalité de la dose est administrée et donc la quantité entrant dans le compartiment est égale à la dose (*i.e.*  $e(0) = Dose$ ). Par la suite, l'évolution de la quantité ne dépend que de la constante de sortie  $k$  (*i.e.* pour  $t > 0$ ,  $e(t) = 0$ ). Pour un médicament administré par voie orale, un compartiment de dépôt fictif, *i.e.* n'ayant pas de volume, est utilisé pour initialiser le système (**Figure 1.6**).



Au temps  $t = 0$ , une fraction  $F$  de la dose est disponible dans ce compartiment et passe du compartiment dépôt au compartiment central, caractérisé par son volume  $V$ , avec une constante d'absorption d'ordre 1  $ka$ . La quantité de médicament dans le compartiment central est cette fois-ci la résultante de la quantité provenant du compartiment dépôt et de

la quantité éliminée selon une constante d'élimination  $k$ , où  $k = \frac{CL}{V}$ . Chaque compartiment est caractérisé par une équation différentielle :

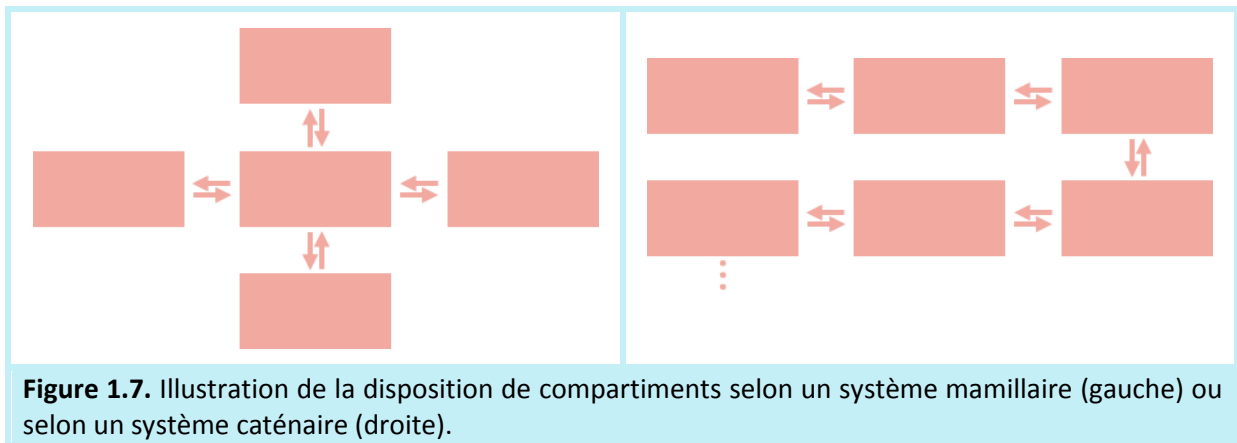
$$\begin{aligned} \frac{dA_1(t)}{dt} &= -ka \times A_1(t), & A_1(t=0) &= F \times Dose \\ \frac{dA_2(t)}{dt} &= ka \times A_1(t) - k \times A_2(t), & A_2(t=0) &= 0 \end{aligned} \quad (1.2)$$

où  $A_1$  correspond à la quantité de médicament dans le compartiment dépôt et  $A_2$  à la quantité dans le compartiment central. La concentration au site de mesure peut être obtenue par  $C(t) = \frac{A_2}{V}$ .

Ce système d'équations différentielles peut être résolu pour déterminer une solution analytique régissant l'ensemble du modèle. Dans cet exemple, une solution analytique du modèle monocompartimental après administration orale est :

$$C(t) = \frac{F \times Dose}{V} \frac{ka}{\left(ka - \frac{CL}{V}\right)} \left( e^{-\frac{CL}{V}t} - e^{-ka t} \right) \quad (1.3)$$

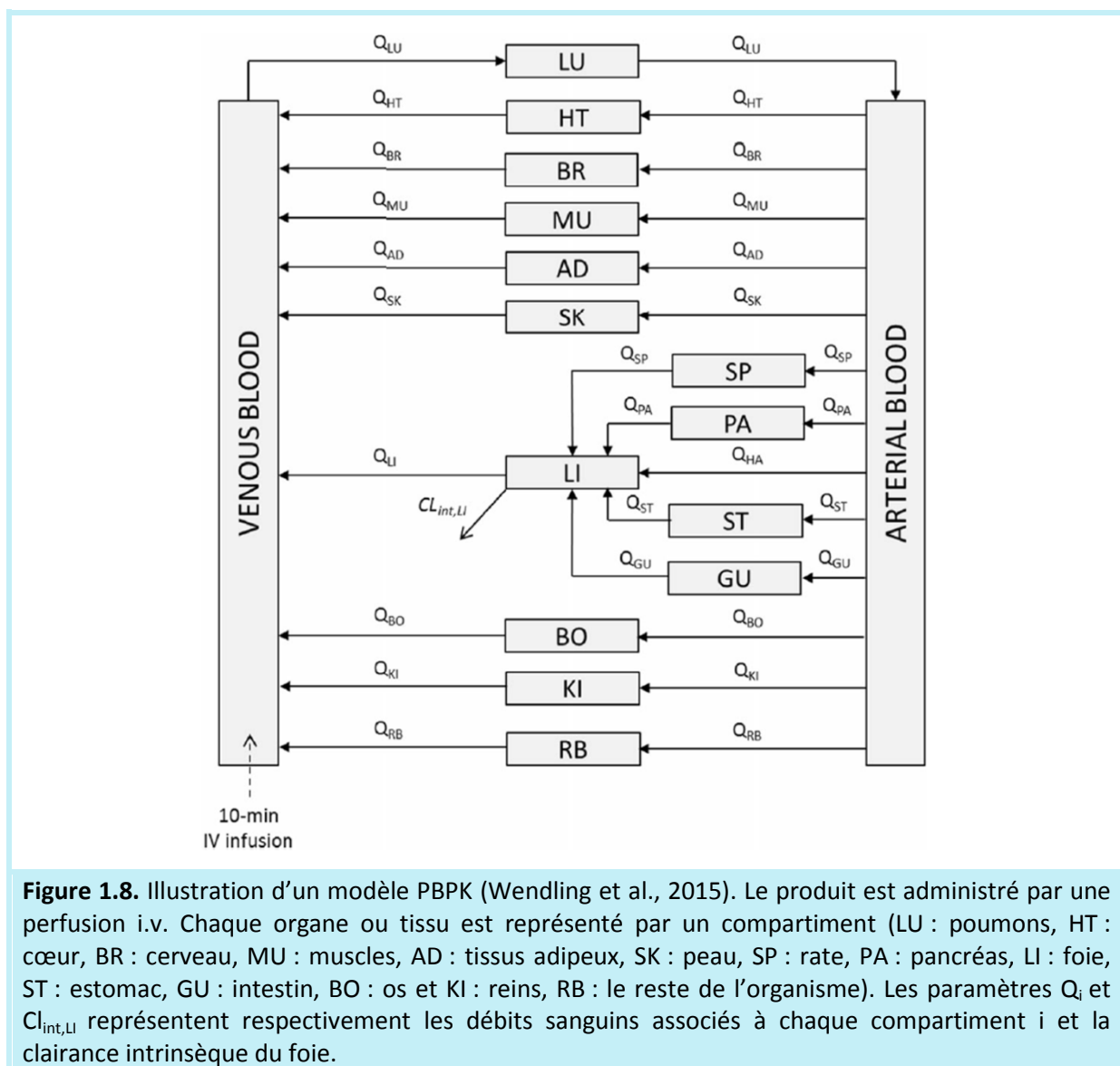
On note  $f(\theta, x)$  le modèle précédent (**équation 1.3**) où  $\theta$  représente l'ensemble des paramètres du modèle à estimer ( $F, ka, V, CL$ ) et  $x$  les variables indépendantes (le temps et la dose). Cette fonction  $f$  est non linéaire car sa dérivée en fonction des paramètres du modèle  $\frac{df}{d\theta}$  dépend de ses paramètres  $\theta$ . On parle de régression non linéaire lorsqu'il s'agit d'estimer les paramètres de ces modèles décrivant au mieux le profil PK d'un individu.



Lorsqu'un modèle comprend plusieurs compartiments, ceux-ci peuvent être disposés selon un système mamillaire où différents compartiments périphériques sont connectés à un même compartiment central, ou selon un système caténaire (dit « en compartiments de

transit ») où les compartiments sont disposés les uns à la suite des autres (**Figure 1.7**). À nouveau chaque compartiment est caractérisé par une équation différentielle.

L'intérêt de cette approche basée sur une régression non linéaire est qu'elle permet d'estimer directement les paramètres physiologiques (volumes, clairances...). Les paramètres secondaires peuvent ensuite être dérivés de l'estimation des paramètres primaires : l'AUC peut être calculée par le ratio de la dose administrée sur la clairance, ou en intégrant la solution analytique (**équation 1.3**) en fonction du temps. De même, connaissant la constante d'élimination  $k$ , la demi-vie peut être calculée de la même manière qu'en NCA.



Une approche plus physiologique existe également et est appelée PBPK (*Physiological Based Pharmacokinetics*). Dans cette approche, les modèles comprennent un grand nombre de

compartiments (**Figure 1.8**), où chaque compartiment représente un tissu ou un organe, reliés entre eux par des débits sanguins (Aarons, 2005). Ainsi les paramètres ont des grandeurs physiologiques. Les volumes des différents compartiments correspondent aux volumes des organes ou des tissus et les constantes de transferts sont souvent exprimées en pourcentage du débit sanguin cardiaque. Ces modèles sont plus complexes et leur développement a été possible grâce à l'amélioration de la puissance de calcul des ordinateurs, permettant d'estimer les paramètres de ces grands systèmes d'équations différentielles. Il est toutefois nécessaire de fixer une partie de ces paramètres, les données disponibles n'étant généralement pas suffisantes pour les identifier (la notion d'identifiabilité sera traitée au paragraphe 1.1.5.4). Certains de ces paramètres peuvent alors être fixés à des valeurs trouvées dans la littérature ou mesurées *in vitro* (propriétés physico-chimique de la molécule, données biologiques d'une population...). Cette approche physiologique permet ainsi l'extrapolation, *e.g.* de l'*in vitro* à l'*in vivo*, ou de l'animal à l'Homme.

#### **1.1.4.3. Estimation de la variabilité pharmacocinétique à partir d'analyses individuelles**

Dans les approches non compartimentale et compartimentale présentées ci-dessus, les données de chaque individu sont analysées séparément et un nombre important de concentrations mesurées au cours du temps est nécessaire à l'estimation des paramètres PK, aussi bien secondaires que primaires. Ces approches sont difficilement utilisables dans le contexte de certaines études où les sujets de par leur condition ne peuvent supporter un nombre important de prélèvements (*e.g.* chez le patient ou en pédiatrie), ou lorsque le protocole ne le permet pas (*e.g.* le nombre de prélèvements prévu dans les études de phase II ou de phase III est très souvent limité).

Comme évoqué dans les paragraphes précédents, les processus PK sont variables d'un individu à l'autre. Ces approches appliquées uniquement au niveau individuel ne permettent pas de caractériser cette variabilité. Il est pourtant nécessaire, lors du développement clinique, de quantifier la variabilité au niveau de la population, afin d'anticiper les variations de la pharmacocinétique d'un médicament. Plusieurs approches permettent d'estimer la variabilité des paramètres PK, se basant sur les analyses non compartimentale ou compartimentale.



Une première approche consiste à traiter les données de plusieurs individus comme s'il s'agissait d'un seul sujet (*naive pooling of data*). L'analyse compartimentale de ce méta-sujet permet de quantifier une variabilité *via* l'erreur résiduelle. Cette approche ne tient pas compte de la corrélation entre les observations d'un même sujet. De plus, elle ne permet pas de distinguer la variabilité intra et interindividuelle, toutes deux diluées dans l'erreur résiduelle (Sheiner, 1984).

Dans une seconde approche en deux étapes, les paramètres de chaque sujet de l'étude sont d'abord estimés séparément par NCA ou régression non linéaire. Puis la variabilité est quantifiée par un résumé statistique en calculant la moyenne et la variance des estimations de tous les sujets (Steimer, 1992). L'estimation se faisant au niveau individuel, à nouveau cette méthode nécessite un nombre important de prélèvements par sujet. De plus, dans cette approche en deux étapes, la variabilité est souvent surestimée, car elle ne distingue pas la variabilité interindividuelle de l'erreur faite sur l'estimation des paramètres individuels.

Il était donc nécessaire de développer une approche permettant notamment l'estimation non biaisée de la variabilité interindividuelle.

### **1.1.5. Les modèles non linéaires à effets mixtes**

Au début des années 1970 s'est développée grâce au professeur Lewis Sheiner, une approche dite « de population » reposant sur l'utilisation de modèles non linéaires à effets mixtes (MNLEM, (Sheiner et al., 1972)). Cette approche de population présume qu'un échantillon de quelques dizaines de sujets, à partir duquel sont estimés des paramètres PK et leur variabilité, permet une extrapolation à l'ensemble de la population. Cette extrapolation n'est pas possible avec les méthodes exposées précédemment, car elles ne permettent pas d'estimer correctement la variabilité (*naive pooling of data*), et/ou elles ne se basent pas sur un modèle (NCA et approche en deux étapes). Les paramètres PK sont dans l'approche de population considérés comme des variables aléatoires, décrites par une distribution dont on estime les paramètres (moyenne et variance). Les paramètres individuels sont alors des réalisations tirées de cette distribution. L'ensemble des observations de tous les sujets d'une étude est analysé en une même étape, ce qui rend l'approche utilisable dans les études où le nombre d'observations par sujet est limité.

### 1.1.5.1. Notation

La concentration  $y_{ij}$  chez le sujet  $i$  ( $i = 1, \dots, N$ ) mesurée au temps  $t_{ij}$  ( $t_{ij} = t_{i1}, \dots, t_{in}$ ), après administration d'une dose  $D_i$ , est définie par la fonction non linéaire  $f$  suivante :

$$y_{ij} = f(\theta_i; D_i, t_{ij}) + \varepsilon_{ij} \quad (1.4)$$

où  $\theta_i$  est le vecteur des paramètres individuels et  $\varepsilon_{ij}$  une erreur résiduelle. La valeur du paramètre  $\theta_i$  pour le sujet  $i$  peut être estimée et diffère de la valeur typique dans la population  $\mu$  (effet fixe) grâce à l'ajout d'un modèle de variabilité interindividuelle et l'estimation d'un effet aléatoire  $\eta_i$ . Cet effet aléatoire est tiré d'une distribution définie *a priori* pour les méthodes paramétriques. Ainsi, on fait l'hypothèse que les effets aléatoires suivent une loi, souvent normale  $\eta_i \sim N(0, \omega^2)$ , où l'écart type  $\omega$  quantifie la variabilité interindividuelle du paramètre. La distribution du paramètre PK, dont sont issues les estimations des valeurs individuelles  $\theta_i$ , dépend du modèle utilisé pour associer l'effet fixe et l'effet aléatoire :

$$\theta_i = \mu + \eta_i \text{ suit une loi normale} \quad (1.5)$$

$$\theta_i = \mu e^{\eta_i} \text{ suit une loi log-normale où } \theta_i > 0 \quad (1.6)$$

En pharmacocinétique, on fait généralement l'hypothèse que les paramètres individuels suivent une loi log-normale en modélisant les effets aléatoires selon l'équation 1.6, l'intérêt étant qu'avec cette distribution les paramètres individuels sont strictement positifs. Dans les méthodes non paramétriques, aucune forme n'est imposée pour la distribution des effets aléatoires (Lai and Shih, 2003; Mallet, 1986).

L'erreur résiduelle  $\varepsilon_{ij}$  quantifie la différence entre la prédiction du modèle et la valeur mesurée chez le sujet  $i$  au temps  $t_{ij}$ . Différents modèles peuvent être utilisés pour décrire la variabilité résiduelle. Nous définissons un modèle général combiné :

$$g(\theta_i; t_{ij}) = \sigma_{inter} + \sigma_{slope} f(\theta_i; t_{ij}) \quad (1.7)$$

où  $\sigma_{inter}$  représente une erreur résiduelle additive et  $\sigma_{slope}$  une erreur résiduelle proportionnelle et  $\varepsilon_{ij} \sim N(0, g(\theta_i; t_{ij})^2)$ .

Les processus PK peuvent varier dans le temps chez un même individu. Cette variabilité intraindividuelle peut être capturée lorsque les sujets d'une étude sont suivis sur une longue période ou que les observations sont faites à différentes visites. Elle est prise en compte par

l'introduction d'un nouveau niveau de variabilité permettant aux paramètres du modèle de fluctuer autour d'une moyenne d'une occasion à l'autre. On peut donc également parler de variabilité interoccasion (Karlsson and Sheiner, 1993). Le vecteur des paramètres individuels s'écrit alors (sous sa forme exponentielle) :

$$\theta_{ik} = \mu e^{\eta_i + \kappa_{ik}} \quad (1.8)$$

où  $\kappa_{ik}$  représente l'écart entre les valeurs du paramètre individuel à différentes occasions  $k$ . On fait souvent l'hypothèse que  $\kappa_{ik} \sim N(0, \delta^2)$ .

En étudiant les relations entre des covariables et les paramètres du modèle, on cherche à expliquer une partie de la variabilité de ces paramètres et à identifier les facteurs de variabilité de l'exposition au médicament, de la relation PK/PD et au final de l'effet d'un traitement. Plusieurs de ces facteurs sont cités au paragraphe 1.1.3 (physiopathologiques, pharmacologiques ou génétiques). Les covariables peuvent être divisées selon leur nature, en variables continues (âge, poids...) ou catégorielles (genre, scores...). Le choix de la relation utilisée dans le modèle pour décrire l'effet de la covariable dépend de sa nature. Par exemple, pour une covariable catégorielle binaire comme le genre (homme *versus* femme), l'effet d'une classe est estimé par rapport à une classe de référence :

$$\theta = \mu \times \beta^{CAT} \quad (1.9)$$

Dans cet exemple, la covariable *CAT* prend la valeur 0 pour la classe de référence (femme) et le paramètre de population  $\theta$  pour cette classe est alors égal à  $\mu$ ; *CAT* prend la valeur 1 pour l'autre classe (homme) dont le paramètre de population  $\theta$  est égal à  $\mu \times \beta$ . Le coefficient d'effet  $\beta$  représente donc la variation de la valeur de population entre hommes et femmes (qui peut être avec cette écriture exprimée en pourcentage).

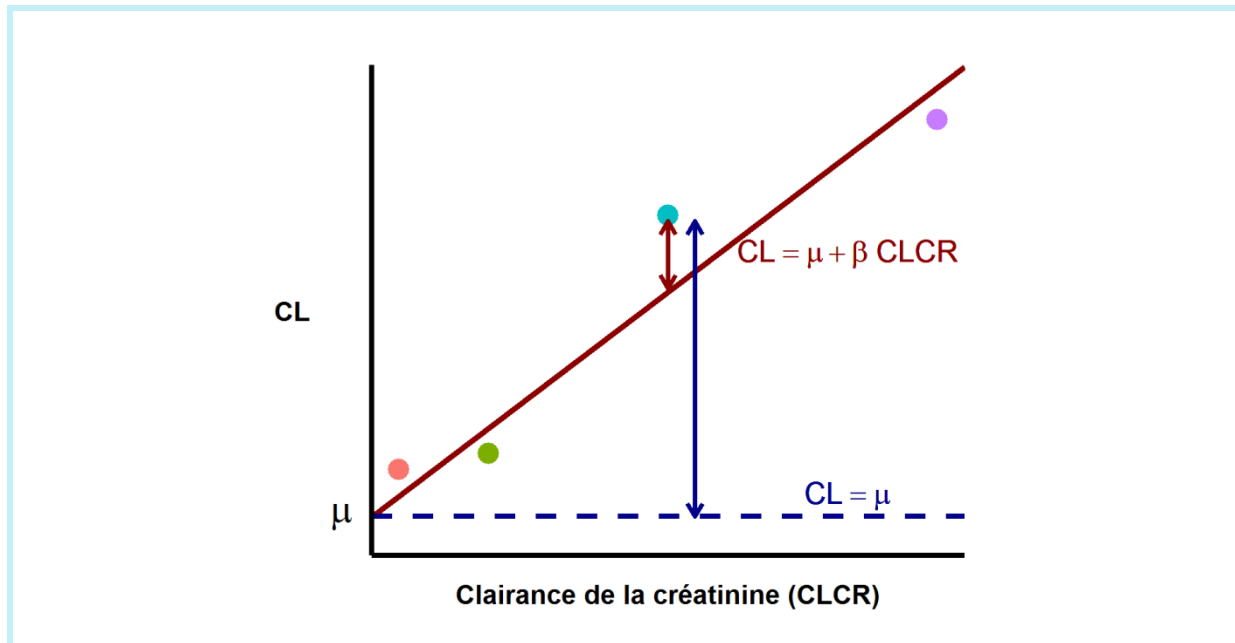
Pour une covariable continue comme le poids, plusieurs modèles peuvent être utilisés comme le modèle linéaire ou le modèle puissance :

$$\theta = \mu + \beta \times (CONT - median(CONT)) \text{ (modèle linéaire)} \quad (1.10)$$

$$\theta = \mu \times \left( \frac{CONT}{median(CONT)} \right)^\beta \text{ (modèle puissance)} \quad (1.11)$$

où *CONT* représente la variable, standardisée par sa médiane. La valeur de l'estimation de  $\beta$  est interprétée différemment selon le modèle choisi.

L'ajout de covariables dans un modèle a généralement comme effet direct de diminuer la variabilité interindividuelle estimée (**Figure 1.9**). Ces covariables peuvent être utilisées pour simuler des profils PK dans des sous-populations (*e.g.* profils chez les femmes *versus* chez les hommes), ou pour prédire la pharmacocinétique d'un médicament chez un sujet en fonction de ses caractéristiques physiopathologiques pour en adapter la posologie.



**Figure 1.9.** Illustration de l'effet d'une covariable sur la variabilité d'un paramètre (adaptée d'un cours de Céline Laffont, « Approche de population utilisant des modèles non-linéaires à effets mixtes »). Les estimations des clairances individuelles sont représentées par les points, en fonction de la clairance de la créatinine des sujets. Le trait bleu pointillé représente la valeur typique de la clairance en absence de covariable (identique pour tous les sujets) et le trait rouge la valeur prenant compte de l'effet de la clairance de la créatinine quantifié par le paramètre  $\beta$ . La variabilité représentée par la flèche bleue en absence de covariable diminue fortement après l'ajout de celle-ci, comme représenté par la flèche rouge.

### 1.1.5.2. Méthodes d'estimation

L'ensemble  $\psi$  des paramètres à estimer dans les MNLEM comprend les effets fixes  $\mu$ , la variance des effets aléatoires  $\omega^2$  de distribution  $D$ , et la variance de l'erreur résiduelle  $\sigma^2$  ( $\psi = \{\mu, \omega^2, \sigma^2\}$ ). L'estimation se fait sur l'ensemble des données (**Figure 1.10**) selon deux approches : une approche fréquentiste basée sur le maximum de vraisemblance, et une approche bayésienne qui utilise une probabilité *a priori* et cherche à maximiser la probabilité des paramètres conditionnellement aux observations pour estimer une distribution *a posteriori*.

Dans l'approche fréquentiste, la vraisemblance des observations  $l(y, \psi) = p(y/\psi)$ , notée  $l$  pour *likelihood*, s'écrit comme le produit des vraisemblances individuelles  $l(y_i, \psi)$  s'obtenant par intégration sur la distribution des effets aléatoires :

$$l(y_i, \psi) = \int_D p(y_i/\theta_i, \psi) d\theta_i \quad (1.12)$$

Dans les méthodes paramétriques, la distribution  $D$  est admise connue et les paramètres de cette distribution sont estimés. L'intégrale présentée dans l'équation 1.12 n'admet pas de solution analytique dans un cadre non linéaire. Des méthodes d'approximation de la vraisemblance sont donc nécessaires pour l'estimation des paramètres dans les MNLEM (Pillai et al., 2005). La première méthode développée se base sur une linéarisation du modèle  $f$  au premier ordre autour des effets fixes (Sheiner et al., 1972). Cette méthode *First Order* (FO) a été développée en 1972 et a abouti à la conception du premier logiciel d'estimation dans les MNLEM, NONMEM ((Beal et al., 2009; Sheiner and Beal, 1980); [www.iconplc.com](http://www.iconplc.com)), qui est aujourd'hui encore le logiciel le plus utilisé dans ce domaine. Une nouvelle méthode de linéarisation a ensuite été développée, la méthode FOCE (*First Order Conditional Estimation*), qui linéarise le modèle autour des effets aléatoires (Lindstrom and Bates, 1990). L'approximation de Laplace proposée par Wolfinger (Wolfinger, 1993) est une extension de ces méthodes et est basée sur une linéarisation du modèle au second ordre. Ces méthodes de linéarisation permettent de se ramener à un modèle linéaire dont la fonction de vraisemblance est explicite. Un algorithme itératif comme l'algorithme de Newton-Raphson est associé à ces méthodes d'approximation de la vraisemblance pour estimer les paramètres de population  $\psi$ . Dans le cas général des modèles non linéaires, des valeurs initiales doivent être fixées pour permettre à l'algorithme de converger vers les estimateurs du maximum de vraisemblance.

D'autres méthodes d'estimation, demandant une plus grande puissance de calcul, se sont développées avec les progrès de l'informatique et permettent le calcul d'une vraisemblance exacte. Ce calcul peut se faire désormais par intégration numérique comme avec la quadrature de Gauss adaptative (Pinheiro and Bates, 2000) ou par intégration stochastique. Cette dernière utilisent des tirages aléatoires de Monte Carlo dans les distributions à intégrer associés à des techniques de chaînes de Markov (Kuhn and Lavielle, 2004). L'algorithme SAEM (*Stochastic Approximation Expectation Maximisation*), développé au début des années 2000 (Kuhn and Lavielle, 2004), utilise cette intégration stochastique et un

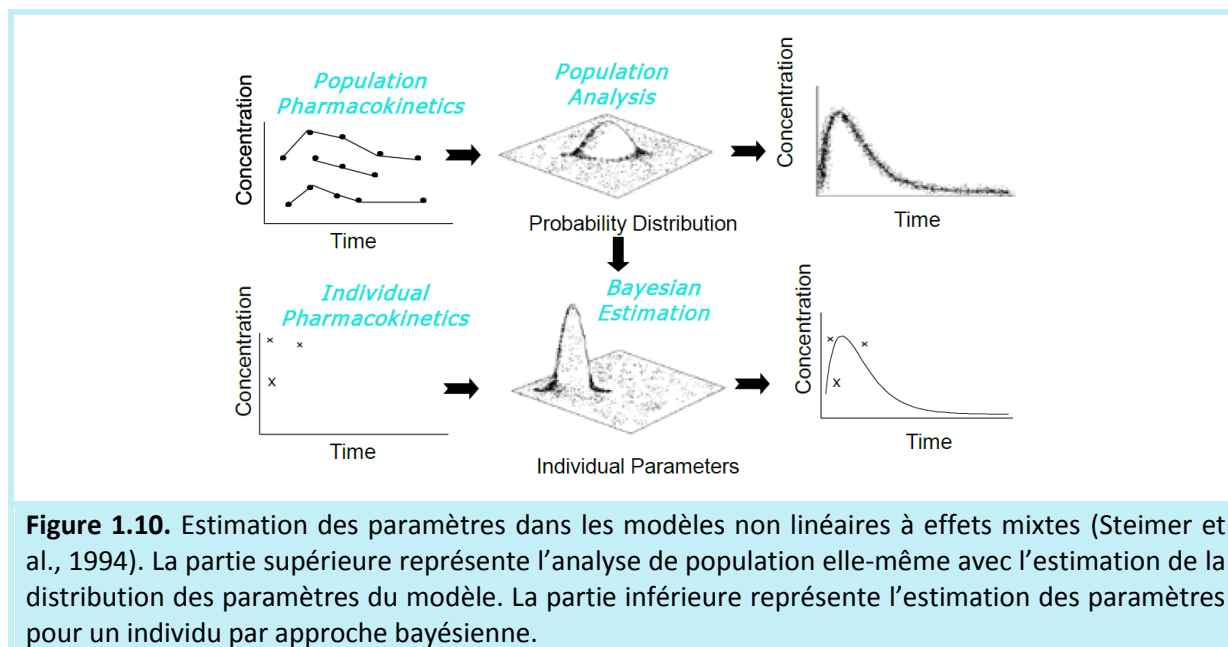
algorithme d'estimation EM. L'étape E (*Expectation*) consiste en une simulation de paramètres individuels avec une méthode de Monte Carlo par chaînes de Markov (MCMC), puis en un calcul de la vraisemblance des données complètes. L'étape M (*Maximisation*) permet une réévaluation des paramètres de population en fonction de la vraisemblance estimée qui sont alors utilisés dans une nouvelle itération et ce jusqu'à convergence de l'algorithme.

Les dernières versions du logiciel NONMEM permettent l'utilisation des algorithmes FO, FOCE, de Laplace ou SAEM, et d'autres logiciels ont également été développés. Ainsi le logiciel R (R Development Core Team, 2012) intègre plusieurs fonctions pour estimer les paramètres dans les MNLEM : la fonction `nlme` utilisant la méthode FOCE, la fonction `nlmer` utilisant l'approximation de Laplace et la fonction `saemix` basée sur la méthode SAEM. Les méthodes d'estimation FO et de Laplace sont également intégrées dans le logiciel SAS (SAS Institute Inc., 2002). Le logiciel Phoenix ([www.certara.com](http://www.certara.com)), conçu pour les analyses NCA ou compartimentales individuelles, intègre aujourd'hui un module d'estimation dans les MNLEM avec la méthode FOCE. Le logiciel Monolix ([www.lixoft.eu](http://www.lixoft.eu)) met en œuvre l'algorithme SAEM et est aujourd'hui un logiciel très utilisé, aussi bien dans le milieu académique que dans l'industrie. Enfin le logiciel NIMROD (pour *normal approximation inference in models with random effects based on ordinary differential equations* (Prague et al., 2013)) permet le calcul d'une vraisemblance exacte par quadrature de Gauss adaptative (Guedj et al., 2007) et intègre un algorithme RVS (pour *robust-variance scoring algorithm*), dérivé de l'algorithme de Newton-Raphson, pour maximiser cette vraisemblance (Commenges et al., 2006). NIMROD permet également l'estimation des paramètres dans les MNLEM par approche bayésienne. L'inférence bayésienne dans les MNLEM et les logiciels associés ne sont pas développés dans ce manuscrit.

### **1.1.5.3. Estimation des paramètres individuels**

Une fois les paramètres de population estimés par les méthodes exposées précédemment, les paramètres individuels peuvent être estimés dans une analyse *post-hoc* (**Figure 1.10**). Les distributions estimées dans l'analyse de population (de variances  $\omega^2$ ) sont utilisées comme lois *a priori*, qui, avec les données mesurées chez le sujet  $i$ , permettent d'estimer des distributions *a posteriori* par une méthode bayésienne (Steimer et al., 1994). Pour un paramètre donné, le mode de la distribution ou *Maximum A Posteriori* (MAP) est généralement utilisé comme estimateur de l'effet aléatoire  $\eta_i$  du sujet  $i$ . Le paramètre PK

individuel  $\theta_i$  est ensuite calculé grâce à la valeur moyenne  $\mu$  et l'effet aléatoire  $\eta_i$  comme dans l'équation 1.6. L'expression *Empirical Bayesian Estimate* (EBE) peut définir aussi bien l'effet aléatoire  $\eta_i$  que l'estimation individuelle  $\theta_i$  du paramètre PK.



L'estimation des paramètres individuels du sujet  $i$  utilise donc une combinaison de l'information *a priori* issue de l'analyse de population et de l'information apportée par les données du sujet.

Plus ces données sont informatives, moins le poids de l'*a priori* sera important dans l'estimation de la distribution *a posteriori* du paramètre. Dans ce cas, cette dernière est centrée autour de la vraie valeur du paramètre. Lorsque le nombre d'observations par sujet est limité on parle de protocole épars. Dans ce cas, le manque d'information apportée par les données donne plus de poids à la distribution *a priori*, la distribution *a posteriori* tendant vers cette dernière. L'effet aléatoire  $\eta_i$  tend alors vers 0 (moyenne de l'*a priori*) et le paramètre  $\theta_i$  tend vers la valeur moyenne  $\mu$ . Ce phénomène de régression vers la moyenne des estimations individuelles est appelé  *$\eta$ -shrinkage*. Le phénomène de régression vers la moyenne (*shrinkage*) a été largement étudié en statistiques, et notamment utilisé pour développer de nouveaux estimateurs (James and Stein, 1961), aussi bien en statistiques fréquentistes que bayésiennes (Efron and Morris, 1977, 1971; Stein, 1956). Plus spécifiquement pour les analyses de population, les conséquences du  *$\eta$ -shrinkage* sur les estimations des paramètres individuels dans les MNLEM ont été mises en évidence dans une

publication de Savic et Karlsson (Savic and Karlsson, 2009). Dans cette publication, les auteurs ont également proposé une métrique quantifiant ce phénomène de  $\eta$ -shrinkage ( $Sh$ ) de la façon suivante :

$$Sh_{\eta}^{sd} = 1 - \frac{sd(\hat{\eta}_i)}{\hat{\omega}} \quad (1.13)$$

où  $sd(\hat{\eta}_i)$  représente l'écart-type de la distribution des estimations individuelles des effets aléatoires et  $\hat{\omega}$  l'écart-type des effets aléatoires estimé dans l'analyse de population (comme expliqué au paragraphe 1.1.5.2). Bertrand et al. (Bertrand et al., 2009) ont par la suite proposé une version de cette métrique basée sur un ratio de variances, plus usuel en statistique (Robert, 1994; Verbeke and Molenberghs, 2009) :

$$Sh_{\eta}^{var} = 1 - \frac{var(\hat{\eta}_i)}{\hat{\omega}^2} \quad (1.14)$$

S'il y a effectivement une régression vers la moyenne, la distribution des estimations individuelles des effets aléatoires se rétrécit autour de sa moyenne, *i.e.* autour de 0, et s'éloigne de la distribution estimée à l'étape de population. Le rapport des deux distributions diminue et le *shrinkage* présenté ci-dessus tend donc vers 1 (soit 100%).

Ce phénomène de *shrinkage* lié à des protocoles peu informatifs affecte également les prédictions du modèle (Savic and Karlsson, 2009). On parle dans ce cas de  $\varepsilon$ -*shrinkage*, qui est associé à une régression des prédictions individuelles du modèle (IPRED) vers les observations (DV). Dans ces travaux de thèse, nous nous intéressons uniquement aux effets associés à  $\eta$ -*shrinkage*.

#### 1.1.5.4. Construction et sélection de modèles

La construction de modèles suivant le principe de parcimonie consiste à trouver le modèle qui décrit le mieux les données observées avec le moins de paramètres. Les modèles dits « sur-paramétrés », en plus d'aller à l'encontre du principe de parcimonie, peuvent être liés à des problèmes d'identifiabilité des paramètres. Un paramètre est identifiable si, à partir des données disponibles, il peut être estimé de façon précise et donc être associé à un estimateur unique (Bellman and Åström, 1970; Godfrey et al., 1994). La difficulté d'accès à certaines données, en matière de faisabilité ou de coûts, oblige souvent à développer des modèles simples en PK/PD. Pour exemple les concentrations dans les compartiments périphériques ne sont généralement pas connues pour le développement de modèles PK et



il est alors nécessaire de fixer certains paramètres pour développer des modèles plus physiologiques, comme en PBPK.

Les critères de sélection et d'évaluation des modèles permettent d'éviter ces sur-paramétrisations.

En pharmacocinétique, il est usuel de tester plusieurs modèles de structure; des modèles à un, deux ou trois compartiments, avec une absorption d'ordre 0 ou d'ordre 1 présentant ou non un temps de latence pour des administrations extravasculaires. Sont également testés plusieurs modèles d'erreur résiduelle (additive, proportionnelle ou combinée) et différents modèles de variabilité. L'ajout d'effets aléatoires est testé successivement sur chaque paramètre du modèle et conservé ou non selon des critères statistiques et graphiques.

Différents critères statistiques peuvent être utilisés pour comparer deux modèles et sélectionner celui le plus en adéquation avec les données. Deux modèles sont dits « emboîtés » si l'un peut être considéré comme un cas particulier de l'autre. Dans ce cas, le test du rapport de vraisemblance (LRT) est généralement utilisé. Sous l'hypothèse nulle, la différence des log-vraisemblances des deux modèles, multipliée par un facteur 2, suit asymptotiquement une loi du  $\chi^2$ . L'hypothèse nulle  $H_0$  (*i.e.* les deux modèles sont équivalents) est testée en comparant la valeur obtenue à la statistique de test du  $\chi^2$  pour un nombre de degrés de liberté égal au nombre de paramètres différenciant les deux modèles. Pour déterminer la significativité d'une covariable, le test de Wald est une alternative au LRT suivant aussi une loi du  $\chi^2$ , et utilise le rapport de l'estimation du coefficient d'effet associé à la covariable sur son erreur d'estimation. Dans le cas de modèles non emboîtés, d'autres critères doivent être utilisés. Les critères d'Akaike (AIC, (Akaike, 1974)) et de Schwartz (BIC, (Schwarz, 1978)) sont les plus utilisés, pondérant la log-vraisemblance par le nombre de paramètres du modèle et, en plus pour le BIC, le nombre d'observations. Pour ces critères, il n'y a pas de test et le modèle présentant le plus faible AIC ou BIC est sélectionné.

Il n'y a pas de consensus sur la méthode de construction d'un modèle. Généralement un premier modèle de base est développé, sans covariable. Puis des covariables sont incluses selon leur impact sur les critères de sélection pour obtenir un modèle final. Le choix des covariables à tester se base avant tout sur des critères biologiques et physiologiques. Il faut en effet être capable d'expliquer le mécanisme liant la covariable au paramètre PK ou PD. Une covariable est conservée dans le modèle tout d'abord selon des critères statistiques (LRT ou test de Wald). Des critères cliniques sont également considérés (*i.e.* la pertinence

clinique de l'amplitude de l'effet de la covariable). De plus, une covariable est généralement conservée si elle explique une partie de la variabilité du paramètre. Mais dans certains cas, une covariable exprimée chez quelques sujets n'explique qu'une faible part de la variabilité, bien que l'effet soit significatif chez ces sujets. Une pré-sélection des covariables est généralement effectuée en amont de l'inclusion dans le modèle de structure en explorant les associations entre les estimations individuelles (EBE) des paramètres du modèle et les covariables mesurées. Des critères statistiques, comme le coefficient de régression de Pearson, sont généralement utilisés pour sélectionner les covariables à tester dans le modèle. Le  $\eta$ -shrinkage expliqué précédemment peut diminuer la puissance de détection de cette étape, comme expliqué dans le paragraphe suivant.

#### **1.1.5.5. Évaluation de modèles**

Comme le montre les paragraphes précédents, de nombreuses hypothèses sont faites lors du développement d'un modèle quant aux choix du modèle de structure, du modèle de variabilité et du modèle de l'erreur résiduelle. Ces hypothèses doivent être évaluées et validées.

Différents critères numériques et graphiques sont utilisés pour l'évaluation. Les autorités de santé ont fourni des recommandations pour le report des résultats et notamment l'évaluation de modèles dans les analyses en pharmacocinétique de population (European Medicines Agency (EMA), 2006). De nombreux outils sont disponibles pour l'évaluation des modèles (Brendel et al., 2007). Nous présenterons ici les critères liés à la précision d'estimation des paramètres, les conséquences du shrinkage des paramètres individuels et évoquerons les principales évaluations graphiques.

L'erreur standard (SE, *Standard Error*) associée à l'estimation de chaque paramètre du modèle quantifie la précision de cette estimation. Le résultat mesurant la précision d'estimation généralement reporté est l'erreur standard relative (RSE, *Relative Standard Error*), calculée comme le ratio de la SE sur l'estimation du paramètre associé. Cette métrique est exprimée en pourcentage. Une faible précision d'estimation, qui se traduit par une SE élevée, est due à un manque d'information, *i.e.* un manque de données, pour l'estimation du paramètre. Cela peut être le signe que le modèle est sur-paramétré, mais également que le protocole de prélèvement de l'étude n'a pas été suffisamment bien conçu pour permettre l'estimation de ce paramètre. C'est pourquoi la précision d'estimation doit être évaluée au regard des données disponibles.

Un fort *shrinkage* des paramètres individuels peut entraîner différentes conséquences. Ainsi des corrélations peuvent être induites ou au contraire masquées lorsque sont explorées les relations paramètre-paramètre, ne reflétant pas la relation réelle. De la même manière les relations apparentes entre paramètres individuels et covariables peuvent être modifiées, influençant l'étape de pré-sélection des covariables. Dans leur publication, Savic et Karlsson ont défini une valeur seuil de  $\eta$ -*shrinkage* à 20-30% (lorsqu'il est calculé par un ratio d'écart types) au-dessus de laquelle le phénomène de régression est trop important pour avoir confiance dans les estimations individuelles (Savic and Karlsson, 2009). Cette valeur, rapportée au ratio de variances, est de 40-60%. Dans le cas de fort  $\eta$ -*shrinkage*, Savic et Karlsson conseillent exclusivement d'utiliser le test de rapport de vraisemblance pour la sélection de covariables, *i.e.* de tester toutes les relations dans le modèle, sans pré-sélection. Dans une publication plus récente, Combes et al. ont montré par simulations que la puissance de détection des covariables par LRT dans le modèle était autant affectée que la puissance des tests de corrélation sur les EBE (Combes et al., 2014).

L'évaluation graphique des modèles comprend des évaluations basées sur les prédictions et sur les résidus du modèle. La première évaluation consiste à s'assurer que le modèle décrit correctement les données observées en comparant ces données aux profils prédits par le modèle pour chaque individu. Des évaluations graphiques plus avancées sont également nécessaires aux étapes clés du développement d'un modèle. Elles se basent sur des simulations faites à partir du modèle final et d'un jeu de données, tels que les VPC (*Visual Predictive Check*, (Bergstrand et al., 2011)) et les NPDE (*Normalised Prediction Distribution Errors*, (Brendel et al., 2006)). On distingue les évaluations internes qui utilisent les mêmes données que celles ayant servi au développement du modèle, et les évaluations externes, plus avancées, utilisant des données différentes. L'évaluation sur des données externes est fortement recommandée si le modèle développé est utilisé pour l'extrapolation à des données non observées.

#### **1.1.6. Rôle de la modélisation en pratique clinique et dans le développement du médicament**

La fonction première de la modélisation est de définir, *via* les estimations des paramètres du modèle, la PK/PD d'une molécule, sa variabilité et les sources de cette variabilité (covariables), afin *in fine* de fournir des recommandations d'utilisation des médicaments aux praticiens (*e.g.* la posologie). Le modèle peut aussi être utilisé pour répondre à différentes

questions et prendre des décisions, ce qui en fait un outil extrêmement intéressant de plus en plus utilisé dans le développement des médicaments (Rowland et al., 2012). Dans cette thèse sont présentées deux problématiques où les MNLEM sont utilisés : l'individualisation des traitements qui fait le lien avec l'étude de la pharmacogénétique et l'optimisation de protocoles utilisée lors des seconds travaux présentés au chapitre 5.

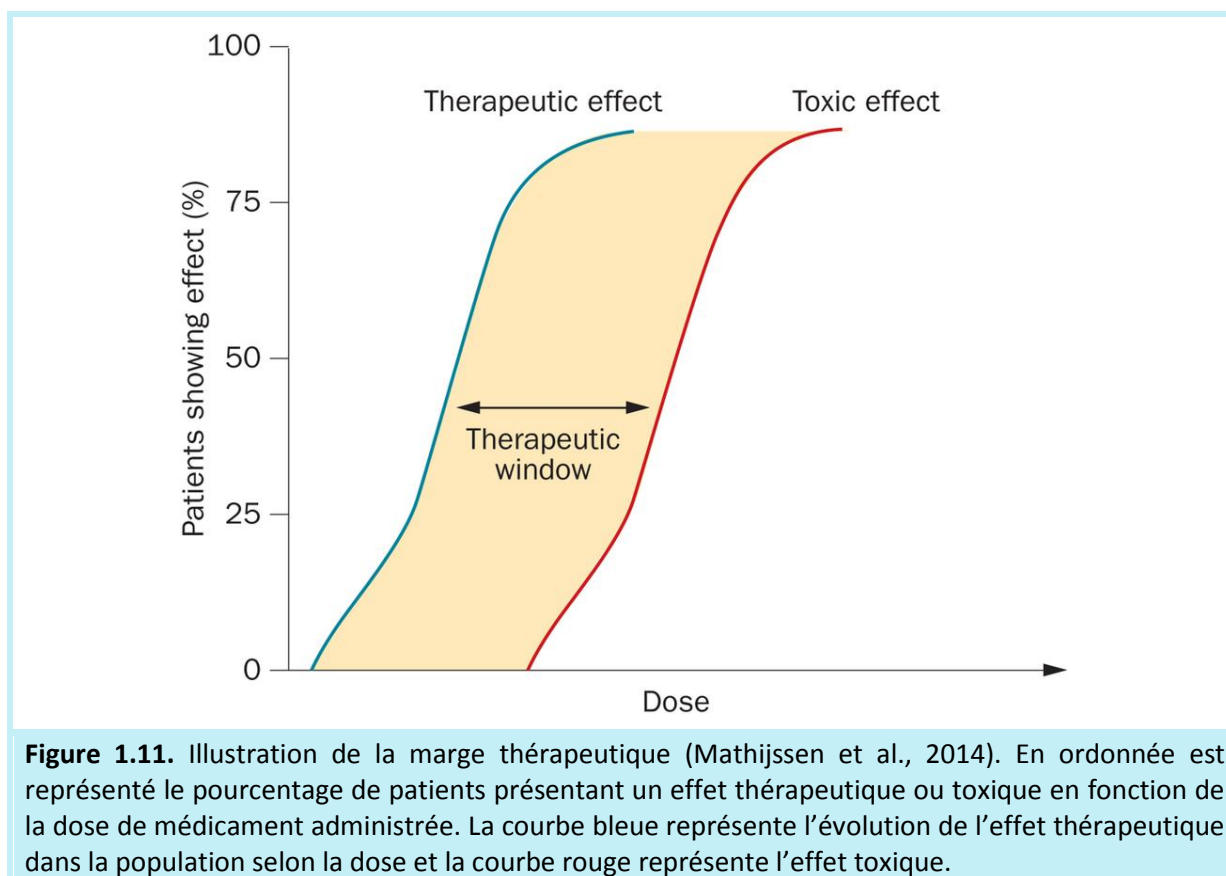
## 1.2. La Pharmacogénétique

### 1.2.1. Définitions

Dès le IV<sup>ème</sup> siècle avant J.C., Pythagore a mis en évidence l'effet de la génétique sur le métabolisme, mettant en garde ses élèves contre des fèves responsables chez certaines personnes de crises hémolytiques aiguës ou « favisme », liées à un déficit enzymatique en Glucose-6-phosphate déshydrogénase (Cappellini and Fiorelli, 2008) (« ce qui peut être nourriture pour certains, peut être poison violent pour d'autres », Pythagore, 510 avant J.C.). Mais le terme « Pharmacogénétique » a été proposé bien plus tard, par Friedrich Vogel (Vogel, 1959), à la suite de travaux d'Arno G. Motulsky mettant en évidence le rôle de la génétique dans l'apparition d'effets indésirables après la prise de médicaments (Motulsky, 1957).

La *Food and Drug Administration* (FDA) définit pour l'industrie la pharmacogénomique (PGx) comme l'étude des variations de l'ADN (acide désoxyribonucléique) et de l'ARN (acide ribonucléique) associées à la variabilité de la réponse aux traitements (Food and Drug Administration (FDA), 2008). Le champ de la pharmacogénomique comprend la pharmacogénétique (PGt), définie comme l'étude de la part de la variabilité entre les sujets dans la réponse aux traitements expliquée par des variations dans la séquence de l'ADN (Food and Drug Administration (FDA), 2008; Weinshilboum, 2003). L'objectif de l'étude de la pharmacogénétique est de pouvoir anticiper par la suite, sachant les caractéristiques génétiques d'un patient, sa réponse à un futur médicament pour en adapter la posologie ou mettre en place un traitement alternatif. Ainsi, la pharmacogénétique est un acteur majeur de la médecine personnalisée (Allorge and Lorient, 2004). Les études pharmacogénétiques visent à mettre en évidence l'impact d'un ou de plusieurs polymorphismes sur la relation dose-effet d'un médicament. Au niveau PK, cela se traduit par une variabilité dans l'exposition de l'organisme au médicament, ce qui modifie la relation concentrations - effet et donc la réponse au traitement. La connaissance des facteurs génétiques pouvant modifier

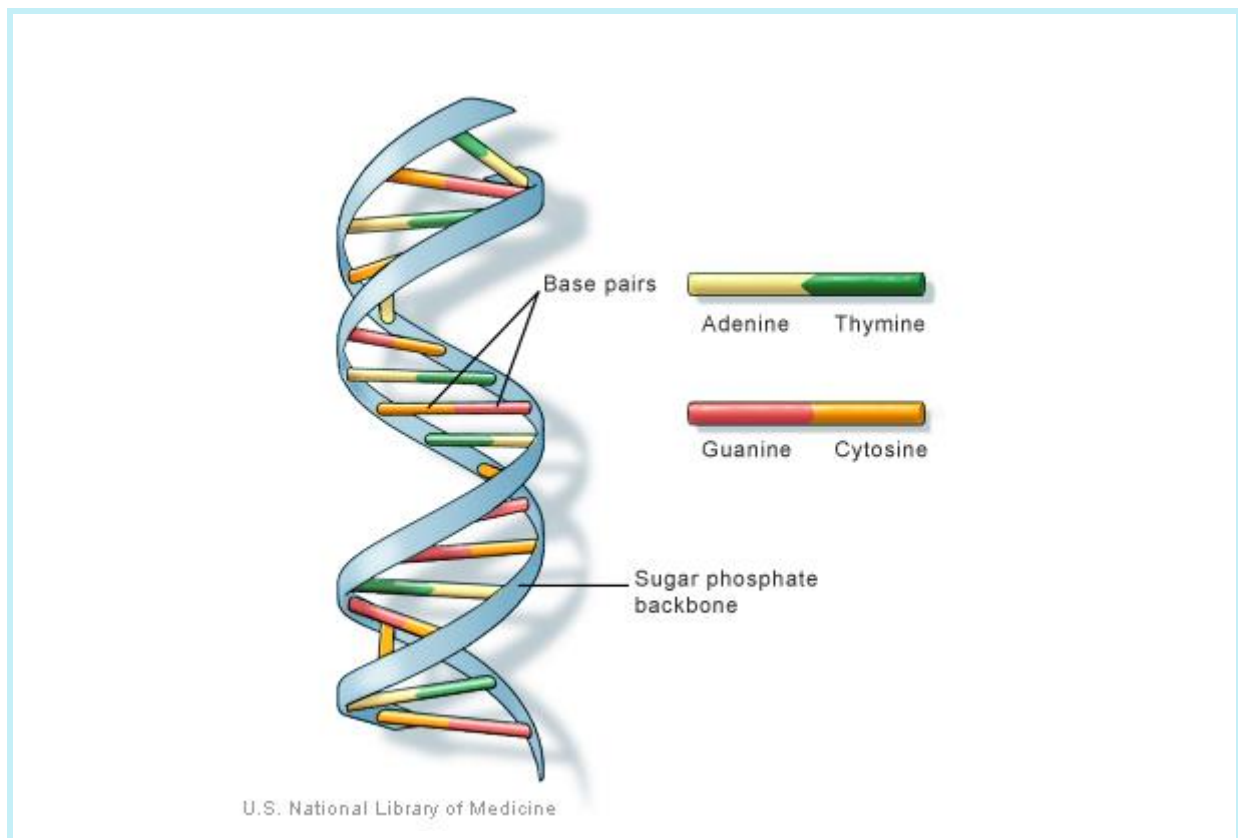
la relation PK/PD d'une molécule est importante pour le suivi du patient et ce d'autant plus que la marge thérapeutique d'un médicament est étroite. Cette marge représente l'intervalle d'exposition au médicament dans lequel le médicament est efficace avec une toxicité limitée. La marge peut être définie en représentant comme illustré dans la **Figure 1.11**, la survenue d'effets thérapeutiques ou toxiques dans une population en fonction de la dose de médicament. Cette relation peut également être définie, pour plus de précision, en fonction de l'exposition du patient au traitement, la relation dose - exposition n'étant pas toujours linéaire. Il est nécessaire de maintenir chez un patient cette exposition et pour cela de connaître les facteurs pouvant interférer et résulter en une exposition trop faible et donc un médicament inefficace, ou une exposition trop importante avec un risque accru de toxicité pour le patient.



### 1.2.2. Les polymorphismes génétiques

Le génome humain contient toutes les informations déterminant les caractéristiques macroscopiques (*e.g.* la couleur des yeux) et microscopiques (*e.g.* le groupe sanguin) d'un individu. Le matériel génétique est codé dans l'ADN, présent dans le noyau des cellules et

réparti sur 23 paires de chromosomes (dont 22 autosomes et une paire de chromosomes sexuels). L'ADN est formé de deux brins en hélice constitués d'un enchainement de nucléotides (**Figure 1.12**). On compte 4 bases azotées, représentées par une lettre, qui constituent le nucléotide : l'adénine (A), la guanine (G), la thymine (T) et la cytosine (C). Les deux brins d'ADN sont reliés au niveau de leurs bases azotées selon des combinaisons connues : l'adénine est toujours liée à la thymine (A-T) et la guanine toujours liée à la cytosine (G-C).



**Figure 1.12.** Structure de l'ADN en double hélice et présentation des quatre bases azotées formant la séquence génétique (US National Library of Medicine). Un brin d'ADN est formé d'un enchainement de nucléotides, eux-mêmes constitués d'un groupe phosphate et d'un sucre, le désoxyribose (*sugar phosphate backbone*), et d'une base azotée où se fait l'appariement avec le brin complémentaire. On parle alors de paires de bases (*base pairs*).

Le génome humain est constitué de plus de 3 milliards de paires de bases, dont environ 1,5% de séquences codant pour des protéines, réparties sur environ 22000 gènes (Lander et al., 2001), dont les protéines impliquées dans la PK/PD des médicaments. Les chromosomes étant présents chez l'Homme par paire, chaque individu possède deux versions d'un même gène.

D'un individu à l'autre, la séquence génétique ne varie que d'environ 0,5%, (Levy et al., 2007) mais ce faible taux de variation est responsable de grandes différences, observables à l'œil nu ou non. À un locus donné (position physique), les différentes versions de la séquence génétique sont appelées « allèles ». Elles sont entre autres responsables de la variabilité observée dans les phénotypes d'intérêt en pharmacogénétique. Ces variations appelées également mutations peuvent être de différente nature. Dans cette thèse, seuls les *Single Nucleotide Polymorphisms* (SNPs) seront abordés, mais d'autres mutations génétiques peuvent être citées pour mémoire : les insertions/délétions d'une partie de la séquence, les mutations chromosomiques comme la trisomie ou la variation du nombre de répétitions d'un motif de plusieurs nucléotides (mini ou microsatellites).

### **1.2.2.1. Les Single Nucleotide Polymorphisms**

Les SNPs représentent l'existence dans la population de variations d'une base dans la séquence d'ADN, *i.e.* l'échange d'un nucléotide par un autre à un locus donné. L'existence dans la population de deux allèles classe les individus en trois catégories : soit aucun allèle n'est muté et l'individu est homozygote sauvage, soit un seul allèle est muté et l'individu est hétérozygote, soit les deux allèles sont mutés et l'individu est homozygote muté. L'information génétique exprimée résulte de l'expression conjointe des deux allèles, ainsi différents modèles génétiques peuvent être observés. On parle de modèle additif lorsque les effets de chaque allèle muté s'additionnent. L'effet sur le phénotype chez un homozygote muté est alors deux fois plus important que chez un hétérozygote. Dans un modèle dominant, la présence d'un seul allèle muté est suffisante pour observer la totalité de l'effet, qui est alors le même chez un homozygote muté et un hétérozygote. À l'inverse, dans un modèle récessif, l'effet n'est observé que si les deux allèles sont mutés et seul le phénotype des homozygotes mutés est affecté.

Cette mutation peut avoir différentes conséquences sur l'expression du gène et l'activité de la protéine synthétisée. Pour les comprendre, il faut expliquer comment l'organisme passe de la séquence génétique d'un gène constituée de plusieurs milliers de bases, à une protéine fonctionnelle. Dans une première, phase l'ADN est transcrit en ARN, une structure monobrin avec une séquence pratiquement identique (la thymine T est remplacée par un uracile U). Cette séquence d'ARN est ensuite traduite en plusieurs acides aminés formant une protéine selon un code génétique connu. Ainsi un ensemble de trois nucléotides ou codon va être traduit en un acide aminé (**Table 1.1**). Ce code est universel car identique pour l'ensemble

du monde vivant (de la bactérie à l'Homme), non ambigu car un codon ne peut donner qu'un seul acide aminé, mais redondant car un même acide aminé peut être obtenu à partir de différents codons. Parmi ces triplets de nucléotides, il existe des codons Stop responsables de l'arrêt de la traduction.

**Table 1.1.** Code génétique : les codons stops sont notés en rouge et le codon d'initiation de la traduction en bleu (Met).

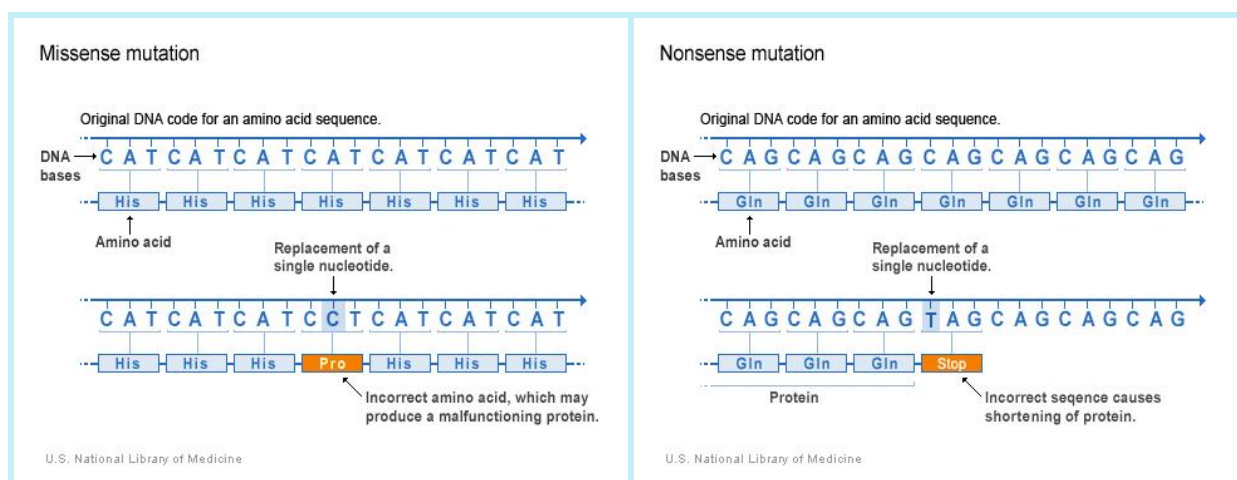
First position	Second position				Third position
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Ala : Alanine, Asp : acide Aspartique, Glu : acide Glutamique, Cys : Cystéine, Phe : Phénylalanine, Gly : Glycine, His : Histidine, Ile : Isoleucine, Lys : Lysine, Leu : Leucine, Met : Méthionine, Asn : Asparagine, Pro : Proline, Gln : Glutamine, Arg : Arginine, Ser : Serine, Thr : Thréonine, Val : Valine, Trp : Tryptophane, Tyr : Tyrosine.

Au locus d'un SNP, l'échange d'une base par une autre peut donner un codon traduit par le même acide aminé; cette mutation est dite « silencieuse » et serait sans impact délétère. Toutefois, il a été démontré que ces mutations silencieuses pouvaient avoir un effet sur l'expression des gènes (*e.g.* sur la vitesse de traduction d'un codon (Chevance et al., 2014)) et donc sur le phénotype (Sauna and Kimchi-Sarfaty, 2011). Il est alors important de ne pas négliger ces mutations dans les analyses pharmacogénétiques. Le codon peut aussi être traduit en un autre acide aminé, la protéine codée ayant une fonction altérée, on parle de mutations « faux sens ». Enfin, le nouveau codon peut être un codon Stop résultant en une



protéine tronquée qui n'a pas d'activité et dans ce cas on parle de mutation « non-sens » (**Figure 1.13**).



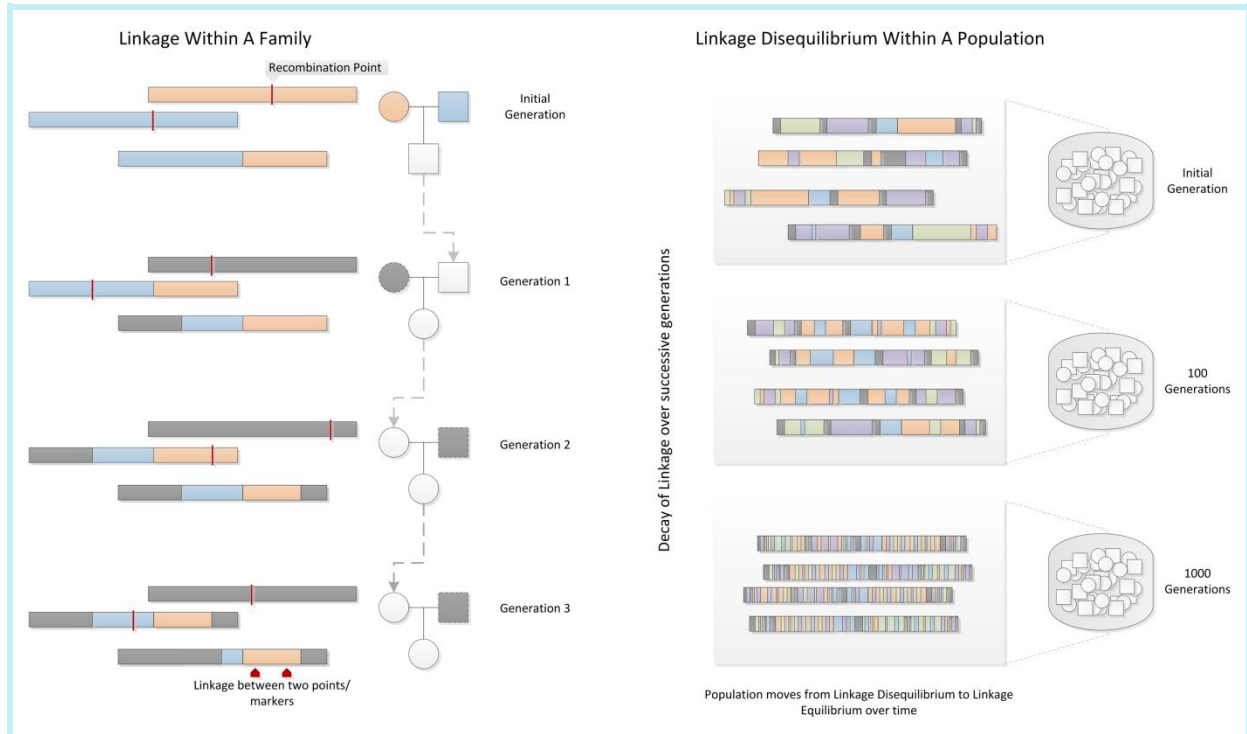
**Figure 1.13.** Illustration d'une mutation faux sens (gauche) et non-sens (droite). Dans la mutation faux sens l'adénine A est remplacée par une cytosine C introduisant le mauvais acide aminé dans la séquence protéique. Dans la mutation non-sens le remplacement de la cytosine C par une thymine T introduit un codon stop résultant en une protéine tronquée.

À ce jour, près de 100 millions de SNPs ont été référencés (d'après dbSNP (*dbSNP Build 144*, résumé du 8 juin 2015), une base de données des variations génétiques hébergée par le NCBI, *National Center for Biotechnology Information* (Wheeler et al., 2007)), ce qui en fait la variation génétique la plus fréquente dans le génome humain, représentant 90% des différences de séquence entre les individus. Les SNPs sont très abondants, distribués dans l'ensemble du génome, et sont donc des marqueurs très informatifs pour l'étude du génome humain. Tous ne sont pas localisés sur les séquences codantes des gènes (exons), mais également sur des séquences non codantes (introns) ou intergéniques où ils pourront affecter la transcription ou l'épissage de l'ARN.

### 1.2.2.2. Le déséquilibre de liaison

Outre leur grand nombre, les SNPs ont une autre caractéristique : ils peuvent être fortement corrélés entre eux. Cette corrélation est due au déséquilibre de liaison (LD pour *linkage disequilibrium*) qui est lui-même la conséquence de recombinaisons génétiques.

La recombinaison génétique consiste en un échange de matériel génétique entre les deux chromosomes d'une paire homologue (**Figure 1.14**). Ces réarrangements permettent un brassage et sont donc à l'origine de la diversité génétique observée dans la population.



**Figure 1.14.** Illustration du déséquilibre de liaison (Bush and Moore, 2012). Sur la gauche, au sein d'une même famille, le déséquilibre de liaison apparaît lorsque deux marqueurs génétiques, représentés par les marques rouges, restent liés sur le même chromosome plutôt que d'être dissociés lors des recombinaisons symbolisées par un trait rouge (points de cassure). Ce phénomène peut être généralisé à une population (à droite) où le poids de la recombinaison est plus important, permettant ainsi le brassage génétique et un retour à l'équilibre de liaison, mais où certains marqueurs restent en déséquilibre.

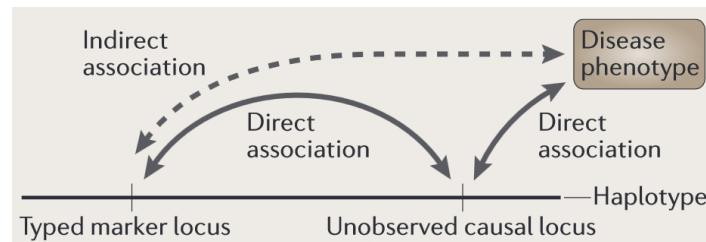
Deux SNPs sont corrélés lorsque la probabilité de recombinaison, *i.e.* que le point de cassure se fasse entre ces deux loci, est faible. Cette probabilité augmente lorsque la distance séparant les deux SNPs augmente. On quantifie cette distance en centiMorgans (cM). Deux SNPs fortement corrélés seront toujours appariés et l'observation d'un seul SNP est suffisante pour avoir l'information sur le second.

Le coefficient de corrélation  $r$  quantifie le déséquilibre de liaison entre deux marqueurs. En prenant l'exemple de deux polymorphismes de génotypes respectifs A/a et B/b, le coefficient de corrélation se calcule comme étant :

$$r = \frac{(p_A \cdot p_B)(p_a \cdot p_b) - (p_A \cdot p_b)(p_a \cdot p_B)}{\sqrt{p_A \cdot p_a \cdot p_B \cdot p_b}} \quad (1.15)$$

où  $p_A$ ,  $p_a$ ,  $p_B$  et  $p_b$  sont les fréquences respectives des allèles A, a et B, b (Lewontin, 1988). Ces deux marqueurs sont en complet déséquilibre de liaison lorsque le carré du coefficient de corrélation, *i.e.* le coefficient de détermination  $r^2$ , vaut 1.

Une conséquence de cet appariement est que l'association mise en évidence lors d'une étude entre un polymorphisme et un phénotype n'est pas systématiquement causative (**Figure 1.15**). Cela peut être une association indirecte et dans ce cas le marqueur observé est en déséquilibre de liaison avec le variant effectivement causal (Balding, 2006). Ceci constitue la base des analyses dites « génome entier » où les variants causaux ne seront pas génotypés, mais identifiés par d'autres variants en déséquilibre de liaison.



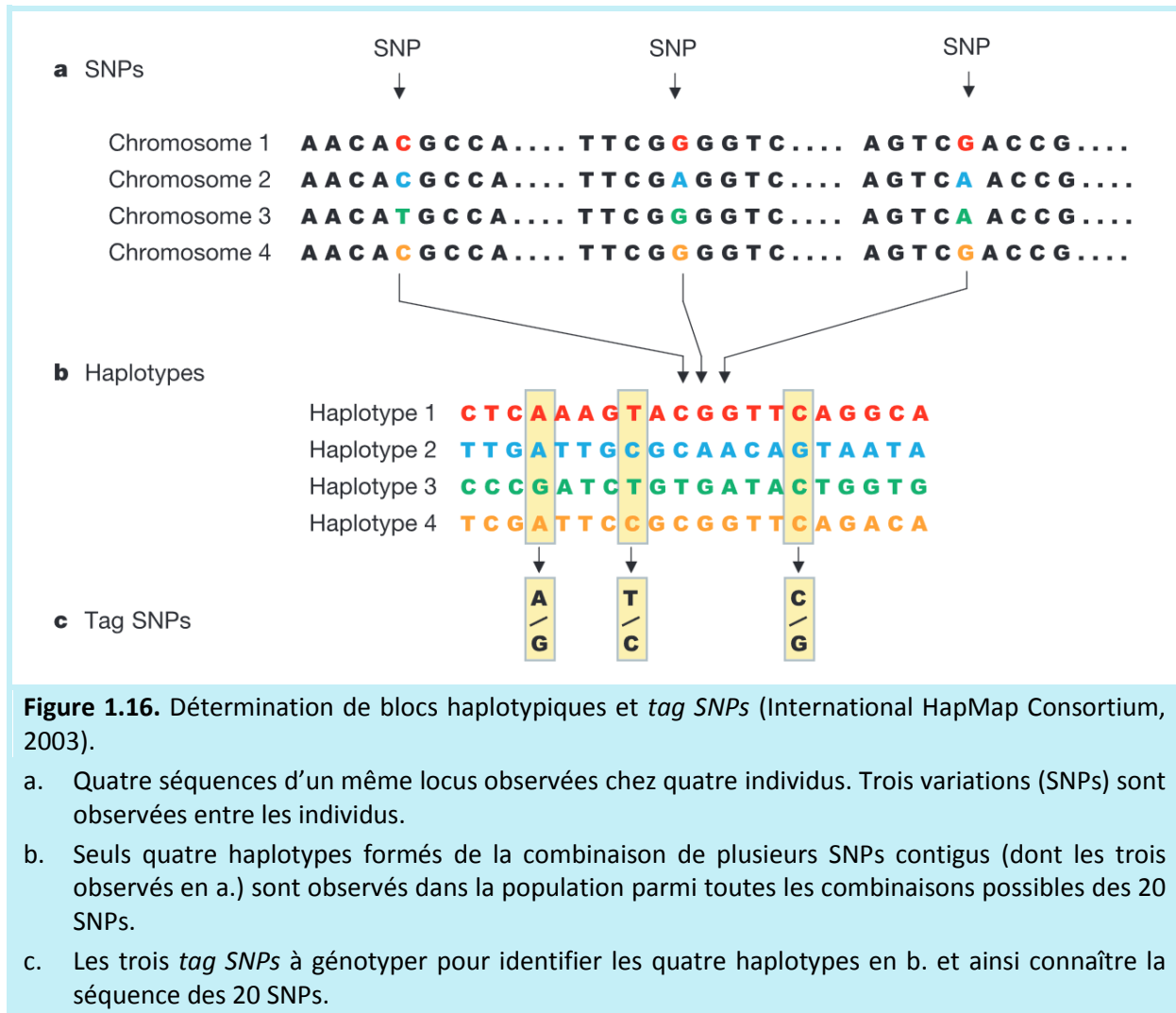
**Figure 1.15.** Conséquences du déséquilibre de liaison (Balding, 2006). Le lien direct entre le phénotype et le variant causal n'a pu être détecté car ce dernier n'a pas été génotypé. Une association a été mise en évidence avec un autre variant qui n'est pas causal, mais en déséquilibre de liaison avec le variant causal.

On peut définir un haplotype représentant un bloc de marqueurs en déséquilibre de liaison se transmettant de génération en génération (**Figures 1.16 et 1.17**).

L'ensemble de SNPs constituant un bloc haplotypique peut être résumé en un sous-ensemble de quelques marqueurs ou « *tag SNPs* ». Ces *tag SNPs* sont considérés comme représentatifs d'une région du génome et leur génotypage suffit à connaître la séquence de toute la région (**Figures 1.16**).

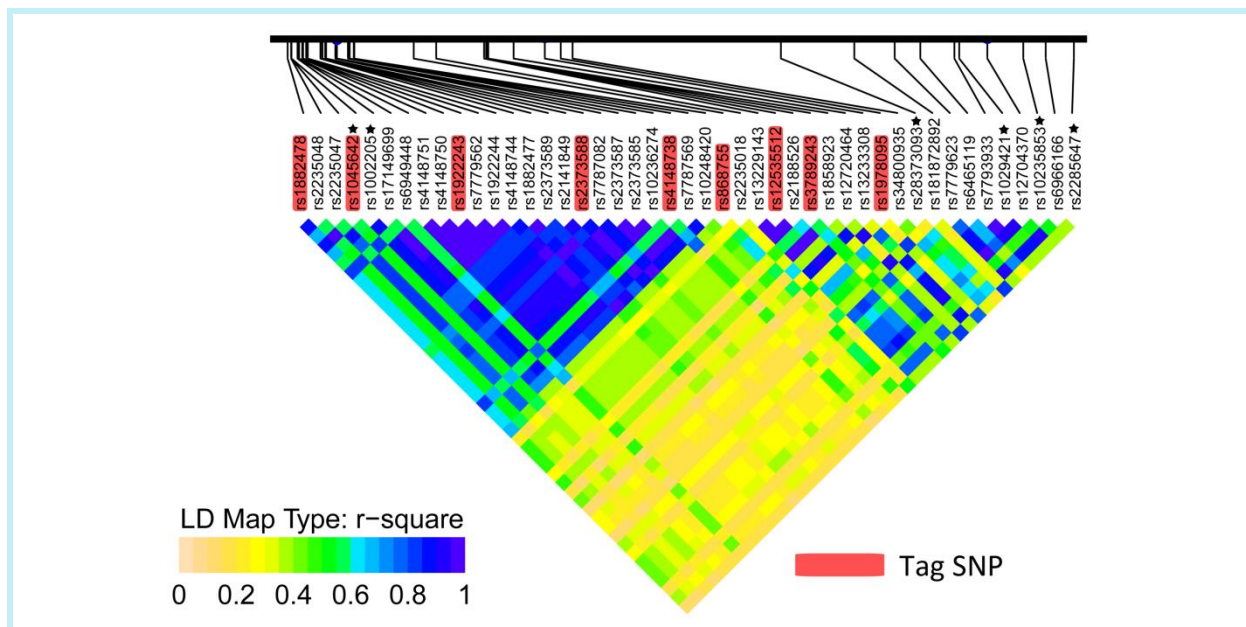
L'analyse en génétique des blocs haplotypiques a comme intérêt de diminuer le nombre de marqueurs à génotyper et le nombre de tests statistiques à effectuer. Nous montrons par la suite que l'exploration d'un nombre important d'associations entre des marqueurs génétiques et un phénotype a un impact sur la puissance des analyses.

Il est tout de même conseillé de commencer une analyse en prenant en compte l'ensemble des polymorphismes (Tregouet and Tiret, 2004). Un polymorphisme peut en effet être retrouvé dans plusieurs blocs haplotypiques, et l'effet détecté pour ce variant est dilué dans les différents blocs et est donc inférieur à l'effet réel du variant pris individuellement.



**Figure 1.16.** Détermination de blocs haplotypiques et *tag SNPs* (International HapMap Consortium, 2003).

- Quatre séquences d'un même locus observées chez quatre individus. Trois variations (SNPs) sont observées entre les individus.
- Seuls quatre haplotypes formés de la combinaison de plusieurs SNPs contigus (dont les trois observés en a.) sont observés dans la population parmi toutes les combinaisons possibles des 20 SNPs.
- Les trois *tag SNPs* à génotyper pour identifier les quatre haplotypes en b. et ainsi connaître la séquence des 20 SNPs.



**Figure 1.17.** Exemple de *LD plot* (Shou et al., 2012). Le graphique représente les corrélations entre chaque combinaison de 2 SNPs. Les zones bleues et violettes présentent les plus forts déséquilibres et en rouge sont surlignés les *tag SNPs*. Dans cet exemple on observe deux blocs haplotypiques.

### ***1.2.2.3. Techniques d'analyses des séquences génétiques : recherche des mutations ponctuelles***

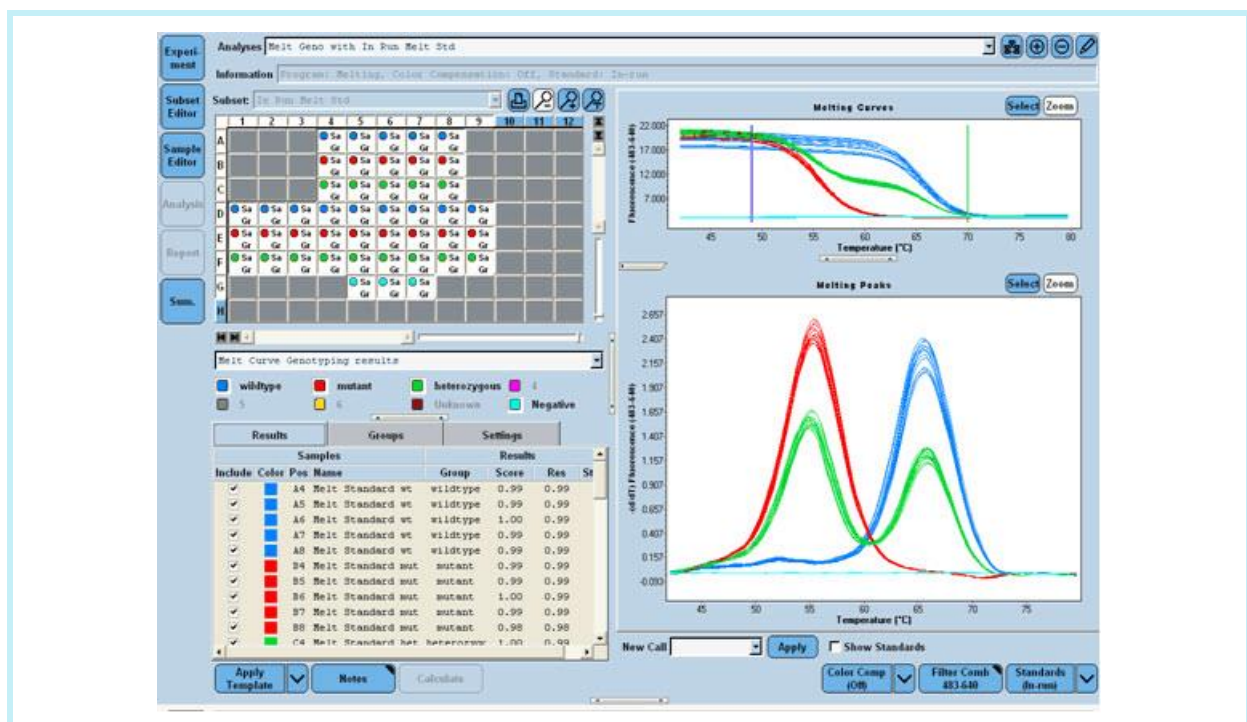
Il existe différentes techniques pour déterminer la séquence génétique d'un individu et identifier les régions de variations. Nous nous intéressons dans cette thèse aux substitutions nucléotidiques (SNPs), des mutations ponctuelles. Nous exposons dans ce paragraphe deux techniques permettant d'identifier ce type de mutation : une technique de séquençage et une technique de génotypage.

La technique de réaction en chaîne par polymérase (PCR, *polymerase chain reaction*) permet d'amplifier en grand nombre des séquences d'ADN à partir d'une faible quantité d'acide nucléique et d'amorces spécifiques. Cette technique s'est développée à partir des années 80 (Mullis et al., 1986) et est à l'origine des méthodes modernes de caractérisation de l'ADN, basées sur le séquençage ou le génotypage. Cette technique permet en effet d'obtenir suffisamment de matériel génétique pour l'analyse.

Le séquençage est une méthode utilisée pour déterminer la séquence génétique exacte d'une portion plus ou moins longue d'ADN. La technique a été établie par Sanger en 1977 (Sanger et al., 1977) et a évolué grâce à l'utilisation de fluorochromes (plutôt que des marqueurs radioactifs) et de séquenceurs automatiques. Un fragment d'ADN monobrin précédemment amplifié par PCR est placé dans un milieu contenant de l'ADN polymérase et des nucléotides (A, T G ou C) dont certains sont modifiés (didéoxynucléotides) et marqués avec des fluorochromes. L'ADN polymérase reconstitue alors la séquence complémentaire de l'ADN monobrin en incorporant de façon aléatoire les nucléotides modifiés, ce qui entraîne un arrêt de l'élongation du brin. La lecture, par des logiciels, de la fluorescence et de la taille des différents fragments formés permet de reconstituer la séquence de l'ADN. En fonction de la région choisie, celle-ci pourra ou non inclure des polymorphismes qui seront alors identifiés. Cette approche ne permet d'étudier que des séquences courtes, de 800 à 1000 bases, et n'est donc pas adaptée à l'étude d'un grand nombre de marqueurs répartis sur l'ensemble du génome. L'émergence de nouvelles méthodes de séquençage (Metzker, 2010) a permis de diminuer le coût de ces techniques et ouvert la voie à des perspectives de marquage plus dense du génome.

Le génotypage est quant à lui un processus permettant de déterminer le variant porté par un individu à une position donnée. La PCR temps réel permet de détecter la présence d'une mutation ponctuelle chez un individu (Lay and Wittwer, 1997). Elle se base sur le principe de

la PCR classique : la séquence d'ADN d'intérêt, contenant la position à géotyper, est amplifiée et hybridée à une sonde complémentaire marquée par fluorescence. La fluorescence ne peut être mesurée que quand le brin d'ADN et la sonde sont hybridés. La force de cette hybridation est diminuée s'il y a eu substitution nucléotidique à la position étudiée, car la séquence de la sonde n'est pas exactement complémentaire. La température du milieu est augmentée progressivement lors d'une phase de fusion après la PCR. Le clivage entre la sonde et l'ADN est atteint à une certaine température, entraînant la diminution de la fluorescence. En présence de substitution nucléotidique, et parce que l'hybridation est moins forte, la température nécessaire au clivage est plus faible. Ainsi on peut identifier si la séquence contient une substitution nucléotidique grâce au profil de fluorescence en fonction de la température (Figure 1.18). Cette approche est relativement abordable et surtout simple à mettre en place grâce à des méthodes automatisées par des robots (e.g. thermocycleurs), mais n'informe que sur la position étudiée.



**Figure 1.18.** Résultats obtenus après génotypages par un thermocycleur (Genotyping using the LightCycler® 480 System, <http://lifescience.roche.com>). Le graphique en bas à droite représente les signaux fluorescents en fonction de la température émis par la sonde. En bleu les individus sont homozygotes sauvages, en rouge homozygotes mutés et en vert hétérozygotes.

C'est avec cette même approche basée sur l'hybridation de brins complémentaires que s'est développée la technique des puces à ADN (Lander, 1999; Lockhart et al., 1996). Dans le cas

des puces à SNPs, des sondes spécifiques à un SNP sont fixées sur une puce en verre. L'avantage ici est que plusieurs centaines à milliers de sondes vont pouvoir être déposées, correspondant à autant de SNPs qui pourront être étudiés. L'ADN du sujet est marqué et mis au contact de la puce. L'hybridation de l'ADN étudié et de la sonde complémentaire pourra alors être mesurée par fluorescence. Ainsi on est aujourd'hui capable de génotyper chez un individu plusieurs centaines de milliers de SNP en une seule expérimentation. Il est possible aujourd'hui de concevoir des puces spécialisées, par exemple dans l'étude de gènes impliqués dans la pharmacocinétique des médicaments, pour des analyses plus ciblées. Cette révolution technologique est à l'origine de la création de projets internationaux de génotypages à grandes échelle, afin de constituer de grandes bases de données génétiques.

#### **1.2.2.4. Les bases de données génétiques**

Il existe aujourd'hui de nombreuses bases de données des variations génétiques qui diffèrent par la manière dont les données sont générées.

Pour mémoire nous citons dbSNP (base de données du NCBI (Wheeler et al., 2007)) ou Ensembl (projet conjoint à l'*European Bioinformatics Institute* et au *Wellcome Trust Sanger Institute* (Hubbard et al., 2007)), qui rassemblent de nombreux marqueurs génétiques, mutations ponctuelles ou autres, de l'Homme mais également d'autres espèces. Ces navigateurs ne génèrent pas eux-mêmes leurs données et sont donc dépendants des laboratoires de recherche qui les produisent. De plus, ils ne sont pas responsables de la qualité des données collectées et ces bases sont régulièrement mises à jour, certains marqueurs étant supprimés car artéfactuels ou bien fusionnés avec d'autres marqueurs déjà présents dans la base.

D'autres projets internationaux génèrent directement leurs données et assurent leur qualité. Nous citons le projet *1000 Genomes* (The 1000 Genomes Project Consortium, 2012) qui regroupe les données issues du séquençage de 1092 génomes humains chez des sujets du monde entier. De la même façon, le projet HapMap ([www.hapmap.org](http://www.hapmap.org), (The International HapMap Consortium, 2005)), débuté en octobre 2002, a pour objectif de cartographier les blocs de SNPs en déséquilibre de liaison et de créer une base de données publique des variants génétiques chez l'Homme. L'intérêt d'identifier ces haplotypes est de déterminer un ensemble de marqueurs (*tag SNPs*) à génotyper afin de connaître des séquences de régions entières. Ainsi, l'identification de plusieurs centaines de milliers de SNP, sans *a priori*, répartis sur l'ensemble du génome, peut permettre l'identification indirecte d'associations



entre un phénotype, généralement des maladies complexes, et un ou des variants causaux. Plusieurs millions de SNPs ont été génotypés dans le projet HapMap, et ce dans différentes populations ethniques.

### **1.2.3. Les analyses pharmacogénétiques**

#### **1.2.3.1. Les analyses de liaison**

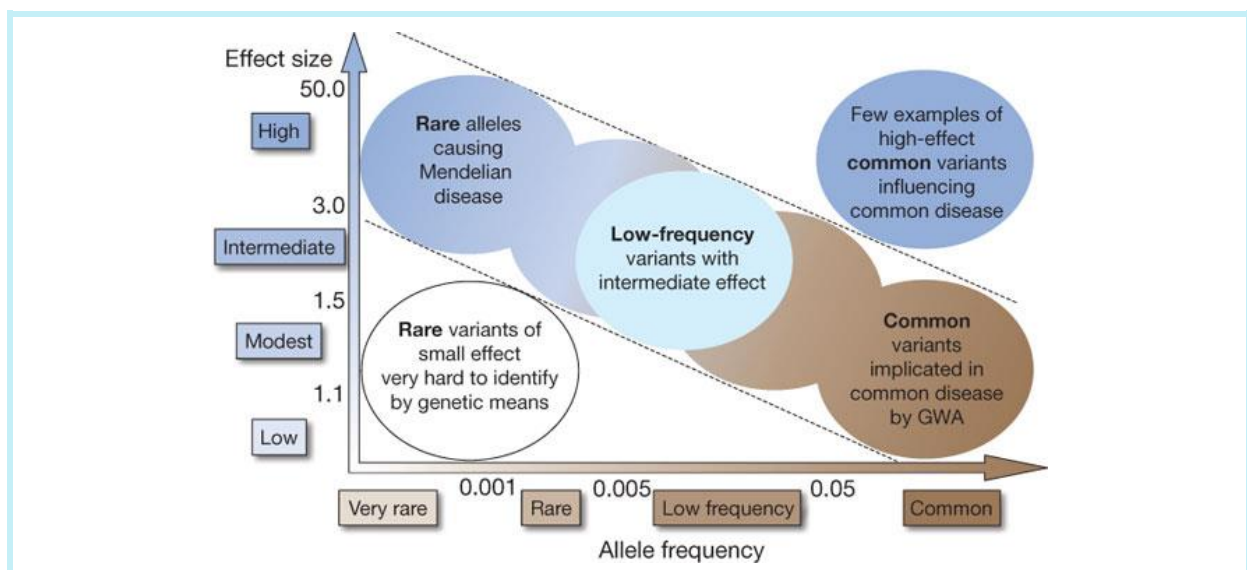
Les maladies monogéniques sont causées par la présence d'une unique mutation dans la séquence génétique d'un individu. On les appelle également maladies mendéliennes, causées par un variant très rare mais associé à un effet important (**Figure 1.19**). Pour identifier le polymorphisme responsable d'une maladie monogénique, une analyse de liaison peut être effectuée dans une famille. On cherche alors une séquence génétique transmise avec la maladie d'une génération à l'autre (Risch and Merikangas, 1996). La région une fois identifiée est séquencée afin de détecter le variant causal. Cette approche n'est pas adaptée aux maladies complexes, causées par de multiples facteurs environnementaux et génétiques. De la même manière, on fait l'hypothèse que la variabilité des phénotypes PK/PD est la résultante de plusieurs facteurs à la fois génétiques et non génétiques. Les études d'association sont plus appropriées car plus puissantes pour ces relations PGt.

#### **1.2.3.2. Les études d'association génétique**

Pour identifier l'effet d'un ou plusieurs variants génétiques sur un phénotype binaire, une étude d'association compare la fréquence de ces variants dans une population divisée en un groupe cas et un groupe témoins, ou en regard de la distribution du phénotype dans la population lorsqu'il est continu. Un marqueur génétique est alors traité comme une covariable catégorielle et on cherche à expliquer une partie de la variance du phénotype, à l'instar d'autres types de covariables comme montré dans le paragraphe 1.1.5.1. La part de la variance expliquée par le polymorphisme dépend de deux facteurs, la fréquence de ce polymorphisme et son effet sur le phénotype (**Figure 1.19**). Un variant fréquent et présentant un effet fort sur le phénotype a tendance à expliquer une part importante de la variance de ce dernier et peut donc être facilement identifiable. À l'opposé, il est difficile d'avoir la puissance suffisante pour mettre en évidence un polymorphisme rare et associé à un effet limité, expliquant une faible part de la variance du phénotype. Mais ces variants sont cliniquement négligeables. Entre ces deux extrêmes, on peut trouver un variant ayant un effet fort mais affectant peu de sujets (cas des maladies monogéniques), à cause de la faible fréquence du polymorphisme dans la population étudiée. Ce variant au niveau de la



population explique une faible part de la variabilité et ne peut être détecté que si la variation est observée chez un nombre suffisant de sujets. Enfin, un variant avec un effet limité mais fréquent explique néanmoins une partie non négligeable de la variance du phénotype dans la population. C'est l'hypothèse du *common variant – common disease* (Pritchard and Cox, 2002), présumant que les maladies complexes, fréquentes, sont causées par des variations génétiques communes dans la population (fréquence supérieure à 5% dans la population), associées à des effets modérés. Ce type de variant est identifié par les études d'association exposées dans ce paragraphe.



**Figure 1.19.** Identification des variants génétiques en fonction de la fréquence de l'allèle mineur représentée en abscisse, et de la taille de l'effet associé à celui-ci représenté en ordonnée et exprimé en odds ratio (Manolio et al., 2009).

La conception de ces études d'association et les résultats qui en sont issus peuvent varier en fonction du mode de sélection des sujets et du nombre de polymorphismes étudiés.

Pour la sélection des sujets d'une étude PGt, une première approche consiste à les sélectionner selon leur génotype, afin d'assurer une bonne représentation des différentes sous-populations génétiques (Comets et al., 2007), notamment des homozygotes mutés dont la fréquence dans la population est plus limitée (Verstuyft et al., 2003). Cette approche garantit une puissance suffisante pour mettre en évidence une relation entre le phénotype et le polymorphisme utilisé comme critère d'inclusion, mais seul un nombre limité de polymorphismes peut être étudié. De plus une hypothèse mécanistique forte est faite sur le lien entre le polymorphisme testé et le phénotype, et on peut alors ne pas détecter d'autres

relations. Les analyses PGt peuvent également être incluses dans des essais plus larges conçus pour d'autres objectifs (*e.g.* étude PK/PD), en génotypant systématiquement les sujets inclus dans l'essai. Elles ont comme avantage de pouvoir étudier un très grand nombre de polymorphismes quand elles sont associées à des méthodes de génotypage comme une puce à ADN. Ici l'étude de la pharmacogénétique se fait sans *a priori*, *i.e.* en aveugle de tout lien entre les polymorphismes et le phénotype d'intérêt. Dans ce cas la distribution des sous-populations ne peut être contrôlée lors de l'inclusion des sujets et est donc dépendante de la fréquence du polymorphisme dans la population. Dans cette thèse nous ne considérons que cette seconde approche avec un génotypage *a posteriori* des études PK.

Le nombre de polymorphismes étudiés permet de diviser les études PGt en trois catégories : les études basées sur une approche « polymorphismes candidats », celles basées sur une approche « gènes candidats » et enfin les études basées sur une approche « génome entier ». Les études polymorphismes et gènes candidats impliquent des gènes sélectionnés *a priori* selon leur fonctionnalité. Dans l'approche polymorphismes candidats un seul gène est étudié et donc seuls un à quelques polymorphismes sont identifiés. Tandis que dans l'approche gènes candidats, plusieurs gènes sont étudiés et quelques dizaines à centaines de polymorphismes présents sur ces gènes sont génotypés. Ainsi ces deux approches s'avèrent être une bonne méthode lorsque le lien mécanistique entre la génétique et le phénotype est déjà connu. C'est notamment le cas en pharmacocinétique où l'on connaît les protéines, et donc les gènes, impliqués dans le métabolisme et le transport des médicaments. Le risque avec cette approche est d'étudier un trop faible nombre de gènes, et donc de ne pas identifier un variant causal. Par ailleurs en raison du nombre de sujets trop limité, les études basées sur une approche gènes candidats manquaient souvent de puissance statistique pour détecter des polymorphismes causaux (Dichgans and Markus, 2005; Zondervan and Cardon, 2007). Mais la bonne connaissance des candidats dans le cadre des études PGt laisse penser qu'elles fourniraient de meilleurs résultats que les études d'association génome entier. Ces analyses génome entier ont vu le jour en 2005 avec la première étude s'intéressant à un grand nombre de variants génétiques dénommée GWAS pour *Genome-Wide Association Studies* (Klein et al., 2005). Cette approche a été rendue possible par le développement des méthodes de génotypages (puces à ADN) et de projets internationaux comme *1000 Genomes* ou HapMap. Du fait du très grand nombre d'associations testées dans ce type

d'étude et donc de la nécessité de corriger le seuil de significativité pour tests multiples, il est difficile de mettre en évidence une relation entre un variant génétique et un polymorphisme. Ces études requièrent un nombre très important de sujets (souvent plusieurs milliers) et nécessitent donc des collaborations internationales (McCarthy et al., 2008). De plus ces GWAS, qui utilisent des puces à ADN et donc des variants fréquents, ne peuvent mettre en évidence que des variants causaux dont la fréquence dans la population est élevée (**Figure 1.19**). Enfin ces études nécessitent la mise en place de nouvelles études confirmatoires (études de réplication) sur des cohortes indépendantes afin d'éviter de conclure à une association faussement positive (McCarthy et al., 2008). Toutefois ces études de réplication devraient être effectuées également pour les études n'entrant pas dans le cadre des GWAS, ce qui n'est à l'heure actuelle pas fait systématiquement.

### **1.2.3.3. Méthodes d'analyses en étude d'association**

L'analyse statistique de l'effet de plusieurs polymorphismes sur un phénotype repose sur l'utilisation de modèles où chaque variant génétique est associé à un coefficient d'effet reflétant son lien avec le phénotype. Dans le cadre des associations PGt, les phénotypes PK/PD étudiés sont généralement continus (*e.g.* la glycémie dans le diabète pour la pharmacodynamie, ou les paramètres PK exposés précédemment), un modèle linéaire est alors utilisé. Les associations entre le phénotype et les polymorphismes génétiques sont quantifiées par régressions linéaires univariées et/ou multivariées. De nombreuses approches ont été développées en génétique sur la base de ces modèles linéaires. Nous exposerons dans ce paragraphe les régressions linéaires classiques où les coefficients d'effets sont d'abord estimés, puis un test est utilisé pour sélectionner les polymorphismes ayant un effet significatif. Puis nous nous intéresserons aux régressions pénalisées où les coefficients d'effets sont estimés et les polymorphismes sélectionnés en une seule étape. Enfin nous parlerons de la correction pour tests multiples, le grand nombre de polymorphismes étudiés avec ces approches étant à l'origine d'une inflation de l'erreur de type I.

De manière usuelle, un modèle linéaire s'écrit sous la forme :

$$Y = a + BX + \epsilon \quad (1.16)$$

où  $Y$  correspond à la variable réponse (le phénotype),  $X$  au vecteur des variables explicatrices (les SNPs),  $a$  l'intercept fixé à 0 pour simplifier le formalisme dans la suite du

texte,  $B$  le vecteur des coefficients d'effet associés aux variables explicatrices et  $\epsilon$  le terme d'erreur, supposée suivre une loi normale. En faisant varier les valeurs prises par les variants  $X$ , il est possible de tester différents modèles génétiques (exposés au paragraphe 1.2.2.1). Ainsi on code les variables  $X$   $\{0, 1, 2\}$  pour un modèle additif,  $\{0, 1, 1\}$  pour un modèle dominant et  $\{0, 0, 1\}$  pour un modèle récessif; respectivement pour un homozygote sauvage, un hétérozygote et un homozygote muté.

Classiquement, on estime les différents coefficients d'effet  $\hat{B}$  en cherchant à minimiser le terme des moindres carrés :

$$\hat{B} = \underset{B}{\operatorname{argmin}} \left( \sum_{i=1}^N \left( y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right) \quad (1.17)$$

où  $y_i$  représente l'observation chez le sujet  $i$  ( $i = 1, \dots, N$ ),  $\beta_j$  le coefficient d'effet associé à la  $j^{\text{ème}}$  ( $j = 1, \dots, p$ ) variable  $x_{ij}$ .

Le principe de l'approche pas à pas, que nous nommons « *stepwise procedure* » dans l'ensemble de la thèse, est d'effectuer une régression univariée sur chaque variable à tester et de ne conserver que la variable la plus significative. Les variables restantes sont de nouveau testées, avec ajustement sur la variable conservée à l'étape précédente, en univarié et de nouveau seule la variable la plus significative est conservée dans le modèle, et ainsi de suite jusqu'à ce que plus aucune variable ne soit significative. Lehr (Lehr et al., 2010) a adapté cette procédure au cas des analyses PGt dans les MNLEM. L'algorithme proposé tient compte à chaque itération du déséquilibre de liaison entre les variants génétiques. Ainsi, après chaque étape de régression univariée, lorsque deux SNPs significatifs sont corrélés ( $r^2 > 0.8$ , **équation 1.15**), seul le plus significatif est conservé dans le modèle.

Les approches basées sur une régression pénalisée sont des méthodes multivariées où tous les coefficients d'effets associés aux variables sont estimés et l'effet de ces variables testé en une seule étape, résultant directement en un modèle final. Nous nous sommes intéressés à trois méthodes : la régression ridge, le Lasso et l'HyperLasso. Dans ces trois méthodes, les estimations des coefficients d'effet sont contraintes par une fonction de pénalisation  $P$ , qui diffère d'une méthode à l'autre :

$$\hat{B} = \underset{B}{\operatorname{argmin}} \left( \sum_{i=1}^N \left( y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right) + P \quad (1.18)$$

La régression *ridge* impose une pénalité sur la taille des coefficients d'effet pour réduire l'erreur de prédiction (Hoerl and Kennard, 1970). Pour cela elle minimise le terme des moindres carrés, pénalisé par la norme  $L_2$ , *i.e.* la somme des carrés des coefficients d'effet :

$$P_R(\delta) = \delta \sum_{j=1}^p \beta_j^2 \quad (1.19)$$

où  $\sum_{j=1}^p \beta_j^2$  est la norme  $L_2$  et  $\delta$  un paramètre de régulation de la pénalisation dont les résultats sont dépendants de la valeur. Ce paramètre  $\delta$  est positif. Lorsqu'il est égal à 0, cela revient à une régression multivariée classique. En augmentant la valeur de  $\delta$ , la pénalisation augmente et les valeurs prises par les coefficients d'effet sont réduites. Pour déterminer la valeur optimale de ce paramètre de régulation en fonction des données de l'analyse, une procédure automatique a été développée dans le logiciel R (Cule and De Iorio, 2013; R Development Core Team, 2012). La régression *ridge* permet d'estimer un modèle en présence de variables fortement corrélées. Cette approche ne permet pas la sélection des variables et l'obtention d'un modèle final parcimonieux, mais elle peut être couplée à une méthode de sélection basée sur une extension du test de Wald (Cule et al., 2011) pour ne sélectionner que les variables significatives.

Le Lasso minimise également le terme des moindres carrés (Tibshirani, 1994), pénalisé cette fois-ci par la norme  $L_1$ , *i.e.* la somme des valeurs absolues des coefficients d'effet :

$$P_L(\xi) = \xi \sum_{j=1}^p |\beta_j| \quad (1.20)$$

où  $\sum_{j=1}^p |\beta_j|$  est la norme  $L_1$  et  $\xi$  un nouveau paramètre de régulation. Plusieurs méthodes, comme la *cross-validation*, peuvent être utilisées pour fixer la valeur du paramètre de régularisation. Nous utilisons dans ces travaux une méthode qui permet de calculer ce paramètre en fonction du nombre de sujets ( $N$ ) et de l'erreur de type I par test  $\alpha$ , en faisant l'hypothèse de conditions asymptotiques :

$$\xi = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{N}{\sigma^2}} \quad (1.21)$$

où  $\Phi^{-1}$  est l'inverse de la fonction de distribution de la loi normale et  $\sigma$  l'erreur standard du phénotype (égale à 1 si le phénotype est standardisé). Le développement de cette formule peut être trouvé dans (Bertrand et al., 2015). L'avantage de cette approche comparée à la régression *ridge*, est que la pénalisation permet de contraindre les effets peu importants à

s'annuler, résultant ainsi en une sélection de variables, sans test, et en l'obtention d'un modèle parcimonieux. Le désavantage de cette méthode est que le nombre de variables sélectionnées dans le modèle est limité par le nombre d'observations, *i.e.* le nombre de sujets (Verzelen, 2012). Cela peut poser problème dans le cas d'études où un très grand nombre de variants génétiques est testé (GWAS). De plus cette méthode a tendance, en présence de variables corrélées, à choisir une variable arbitrairement et à estimer l'effet associé aux autres variants à 0. Un algorithme de sélection de covariables dans les MNLEM basé sur la méthode Lasso est implémenté dans le programme PsN (Lindbom et al., 2005, 2004; Ribbing et al., 2007).

L'HyperLasso est une méthode récente (Hoggart et al., 2008) dérivée du Lasso. Elle utilise une autre forme de pénalisation dépendant de deux paramètres, un paramètre de forme  $\lambda$  et un paramètre d'échelle  $\gamma$  :

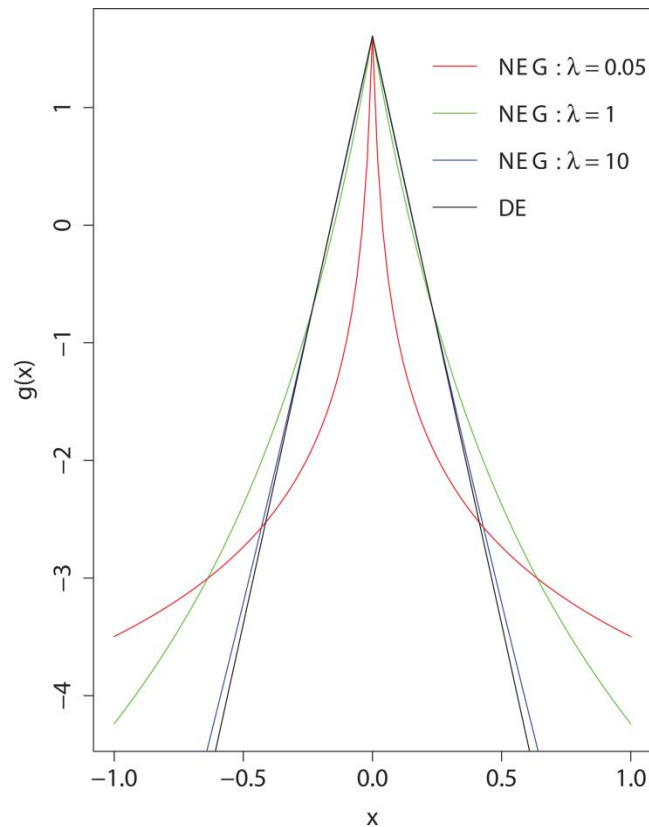
$$P_H(\lambda, \gamma) = - \sum_{k=1}^p \left[ \frac{\beta_j^2}{4\gamma^2} + \log D_{(-2\lambda-1)(\frac{|\beta_j|}{\gamma})} \right] \quad (1.22)$$

où  $D$  est une fonction parabolique cylindrique (Gradshteyn and Ryzik, 1980). Pour déterminer les valeurs des paramètres de pénalisation  $\lambda$  et  $\gamma$ , il est nécessaire de fixer un des paramètres, usuellement  $\lambda$ , pour calculer l'autre. Il a été démontré que fixer le paramètre de forme  $\lambda$  à 1 permettait d'estimer des distributions de coefficients d'effet plus réalistes (Vignal et al., 2011). Le paramètre d'échelle  $\gamma$  peut alors être calculé selon la même méthode que le Lasso, en adaptant la formule (Bertrand et al., 2015; Hoggart et al., 2008) :

$$\frac{\text{sign}(\beta_j)(2\lambda + 1)}{\gamma} \frac{D_{-(2\lambda+2)(\frac{|\beta_j|}{\gamma})}}{D_{-(2\lambda+1)(\frac{|\beta_j|}{\gamma})}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{N}{\sigma^2}} \quad (1.23)$$

La pénalisation utilisée dans ces approches peut être considérée, d'un point de vue bayésien, comme une loi *a priori* appliquée sur la distribution des coefficients d'effet. La régression *ridge* utilise un *a priori* Gaussien qui ne permet pas d'annuler des coefficients. Le Lasso utilise une distribution *a priori* Double Exponentielle (DE) qui pénalise tous les coefficients avec la même amplitude, annulant les effets les plus faibles (**Figure 1.20**). L'HyperLasso utilise une distribution *a priori* Normal Exponentiel Gamma (NEG) qui a l'avantage de présenter un pic resserré autour de 0, favorisant l'élimination des variables ayant un effet faible et donc un modèle parcimonieux, et une queue de distribution large. Cette loi permet de pénaliser moins sévèrement les effets les plus importants (**Figure 1.20**).

Ainsi plus le paramètre de forme sera faible et plus le pic de la distribution sera resserré autour de 0, résultant en la sélection de peu de SNPs corrélés. L'HyperLasso semble montrer un avantage dans les études où un très grand nombre de variants génétiques est étudié.



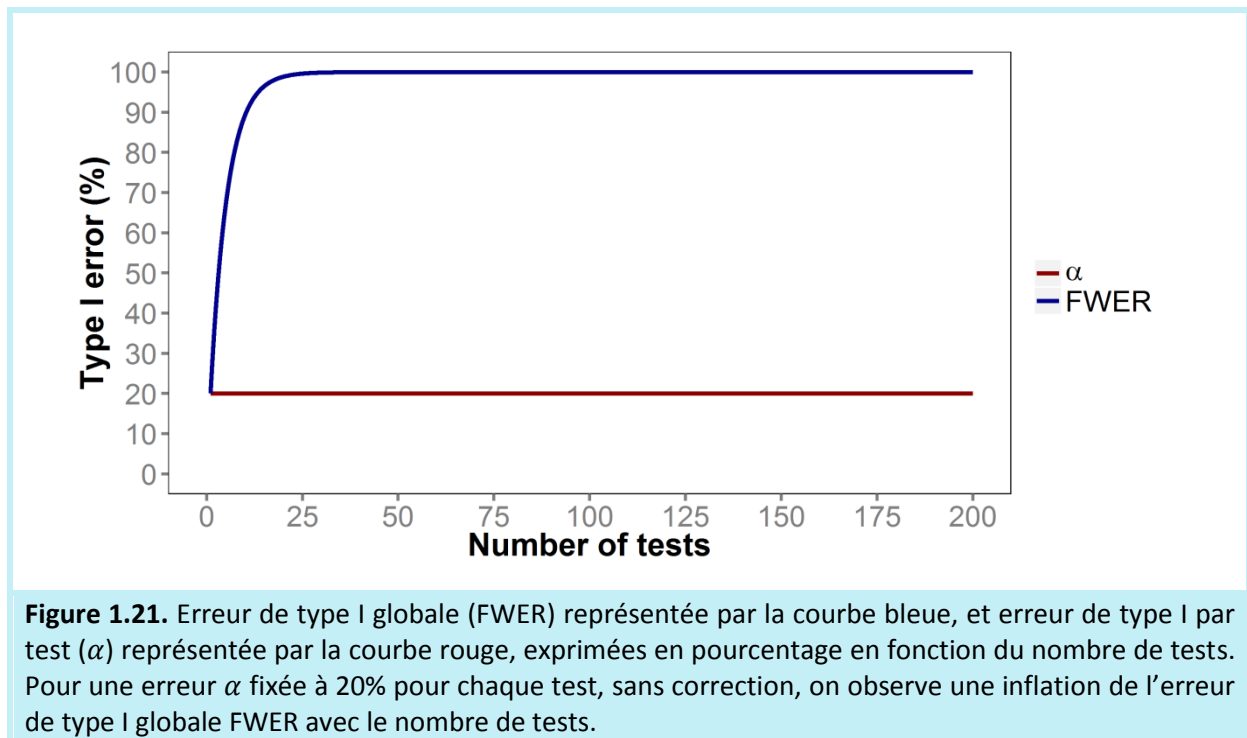
**Figure 1.20.** Distribution *a priori* des méthodes Lasso (DE) et HyperLasso (NEG) en fonction du paramètre de forme  $\lambda$  (Hoggart et al., 2008). La méthode Lasso avec son *a priori* DE représenté en noir pénalise tous les variants avec la même intensité, tandis que l'*a priori* NEG de la méthode HyperLasso représenté en rouge, vert et bleu pénalise plus fortement les faibles coefficients. Plus le paramètre de forme de la distribution NEG est faible et plus cette dernière est resserrée autour de 0 avec une queue large comme représenté en rouge, à l'inverse avec un paramètre de forme très élevé, la distribution NEG représentée en bleu tend vers la distribution DE représentée en noir, et donc la méthode HyperLasso tend vers la méthode Lasso.

Nous l'évoquions précédemment, l'analyse de données génétiques peut amener à tester un grand nombre d'associations sur un même phénotype. Cette multiplicité des tests conduit à une inflation de l'erreur de type I, correspondant à la probabilité de conclure à une association significative alors que celle-ci ne l'est pas. On fixe généralement cette erreur de type I à 5% mais on peut également choisir une probabilité plus faible (*e.g.* 1%) lorsque l'on veut s'assurer de la significativité de l'association mise en évidence, ou au contraire une probabilité plus importante (*e.g.* 20%) dans le cas d'analyses exploratoires où l'on cherche à

générer des hypothèses. Cette erreur de 5% correspond à un seul test. Le *Family Wise Error Rate* (FWER) est la probabilité de détecter au moins un faux positif, *i.e.* une association faussement significative, lorsque plusieurs tests sont effectués. Cette erreur de type I globale est alors égale à :

$$FWER = 1 - (1 - \alpha)^{N_t} \quad (1.24)$$

où  $\alpha$  représente l'erreur de type I pour un test et  $N_t$  le nombre de tests. Dans ce cas le risque global de conclure à tort à une association significative augmente (Huque and Sankoh, 1997), comme représenté dans la **Figure 1.21**.



Le principe des analyses statistiques fréquentistes est de conclure à l'association entre une variable d'intérêt et une variable descriptive si la valeur  $p$  (ou *p-value*), *i.e.* la probabilité que l'hypothèse nulle selon laquelle il n'y a pas d'association soit vérifiée, est inférieure à un seuil égal à l'erreur  $\alpha$ . Plusieurs méthodes de correction pour tests multiples ont été développées. Nous évoquons dans cette thèse deux méthodes basées sur une correction de la valeur  $p$ . La méthode de Bonferroni (Bland and Altman, 1995) consiste à réaliser les tests avec un seuil de significativité corrigé pour qu'au final l'erreur au niveau global FWER atteigne la valeur souhaitée (*e.g.* 20%).



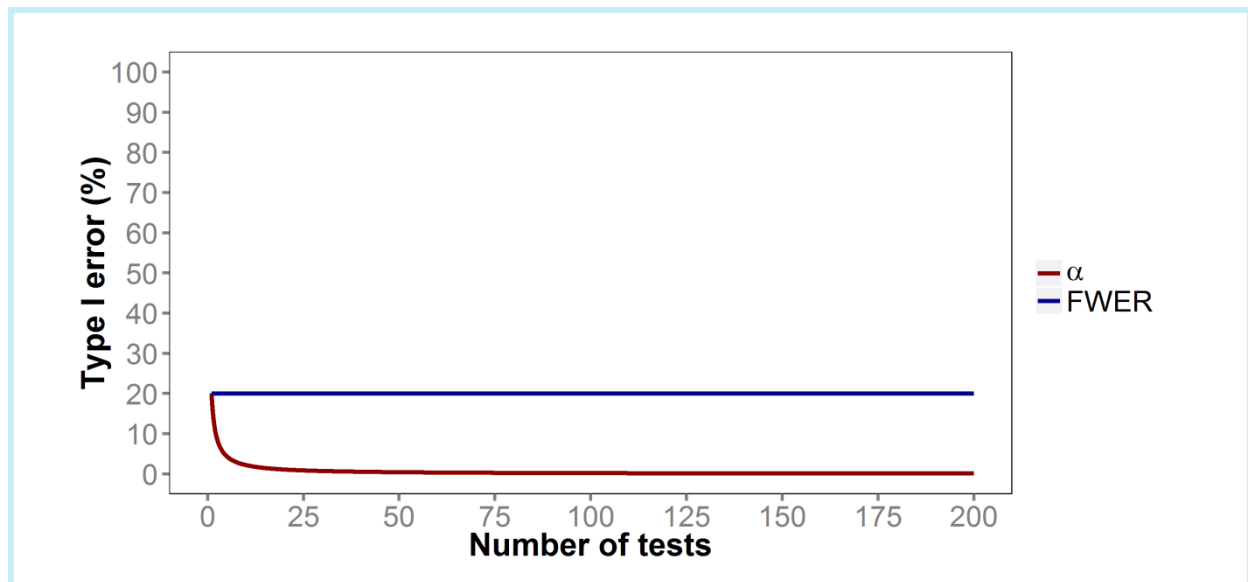
En supposant que les tests sont indépendants, l'erreur de type I par test corrigée par la méthode de Bonferroni est la suivante :

$$\alpha_{Bonferroni} = \frac{FWER}{N_t} \quad (1.25)$$

L'autre méthode est celle de Šidák (Šidák, 1967) où, sur le même principe, on calcule l'erreur de type I par test en fonction du nombre de tests à effectuer d'après la formule (**Figure 1.22**) :

$$\alpha_{\text{Šidák}} = 1 - (1 - FWER)^{1/N_t} \quad (1.26)$$

Ces deux méthodes donnent des seuils très proches pour un faible nombre de tests. Lorsque ce nombre est plus élevé, la méthode de Bonferroni est plus conservatrice.



**Figure 1.22.** Erreur de type I globale (FWER) représentée par la courbe bleue, et erreur de type I par test ( $\alpha$ ) représentée par la courbe rouge, exprimées en pourcentage en fonction du nombre de tests. Pour une erreur globale FWER cible de 20%, l'erreur de type I par test  $\alpha$ , corrigée avec la méthode de Šidák, diminue avec le nombre de tests.

On comprend ainsi le manque de puissance qui peut être observé dans les études en génétique et notamment dans les GWAS. La méthode de Bonferroni est la plus utilisée et conduit à des seuils de significativité très bas ( $< 5 \cdot 10^{-8}$ ) du fait du nombre extrêmement important d'associations testées. Ainsi, seules des associations avec un effet très important et incluant des variants très fréquents (**Figure 1.19**), sont mises en évidence dans ces études. Pour atteindre la puissance nécessaire à la mise en évidence de relations moins importantes,

il faudrait inclure dans ces études des cohortes de plusieurs milliers de sujets, ce qui est difficilement réalisable en pratique dans des études PK/PD (Hong and Park, 2012).

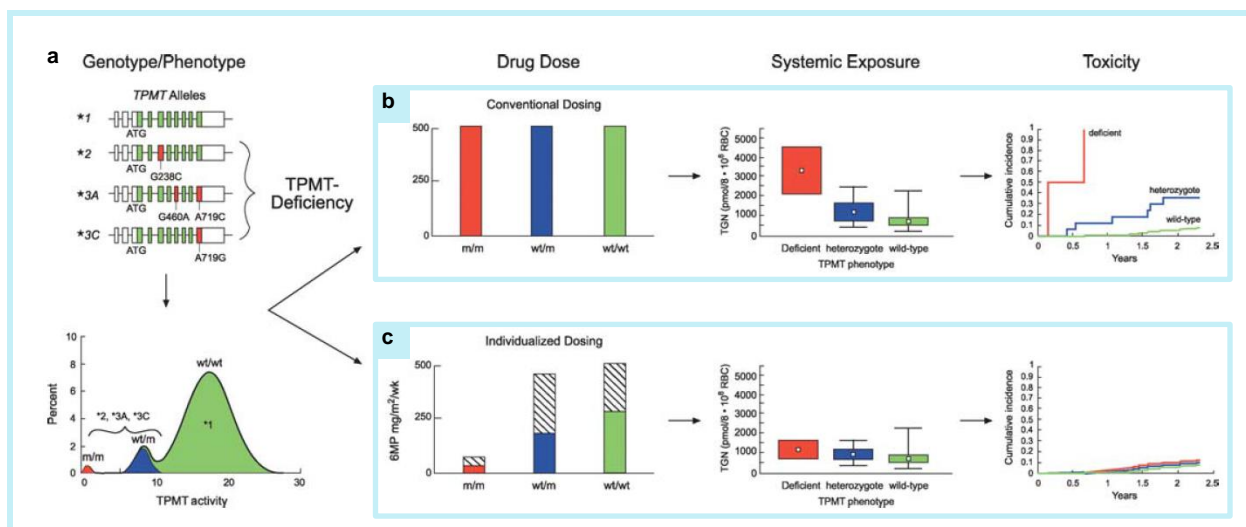
#### **1.2.4. La pharmacogénétique en pratique clinique**

L'intérêt d'étudier la pharmacogénétique des médicaments lors de son développement clinique, ou après sa mise sur le marché, est d'identifier des facteurs génétiques liés à la variabilité dans la réponse aux traitements, aboutissant à des recommandations d'usage. La génétique d'un patient est alors un facteur à prendre en compte dans l'individualisation du traitement. Les modèles PK/PD sont un des outils d'aide à la prise de décision dans le cadre de la médecine personnalisée et en fonction des données disponibles pour le patient, deux approches d'individualisation du traitement sont envisageables : une adaptation *a priori* et une adaptation *a posteriori* par une approche bayésienne.

L'adaptation *a posteriori* reprend la même méthodologie que celle utilisée pour l'estimation des paramètres individuels après une analyse de population (expliquée au paragraphe 1.1.5.3). Cette approche nécessite d'avoir administré le médicament au patient pour estimer les paramètres PK individuels par une méthode bayésienne à l'aide des distributions *a priori* du MNLEM et des concentrations mesurées après administration. Grâce aux paramètres individuels, les concentrations suivantes peuvent être prédites afin de réajuster la dose pour obtenir l'exposition et l'effet souhaité. L'adaptation *a priori* est utilisée en amont de la mise en place du traitement, alors qu'aucune information sur les concentrations du médicament chez le patient n'est disponible. Les caractéristiques physiologiques et/ou biologiques sont utilisées pour imputer les covariables présentes dans le modèle afin de prédire les valeurs des paramètres du modèle propres à l'individu. Par exemple un modèle décrivant la pharmacocinétique du carboplatine, un anticancéreux, a été développé et démontrait l'association entre la clairance du produit et quatre covariables (âge, poids, sexe, créatininémie). Ainsi, à partir des données du patient pour ces quatre variables, la clairance du carboplatine peut être prédite et la dose optimale pour atteindre l'exposition souhaitée calculée comme étant  $Dose = AUC \times CL$ , où AUC prend une valeur cible (Chatelut et al., 1995). Cette approche a l'avantage d'être effectuée avant toute administration du traitement, ainsi le patient n'est pas exposé à des doses inappropriées (inefficaces ou toxiques). La pharmacogénétique s'inscrit dans cette approche *a priori*, le traitement étant ajusté grâce à l'information sur les covariables génétiques du patient.

De nombreuses analyses pharmacogénétiques ont été publiées ces dernières années, certaines associations se traduisant au niveau clinique par des recommandations d'usage (Becquemont et al., 2011; Sim and Ingelman-Sundberg, 2011).

Un premier exemple marquant associant des polymorphismes génétiques à un risque de toxicité concerne l'abacavir, un traitement utilisé dans l'infection au VIH et induisant des phénomènes d'hypersensibilité chez 5% des patients traités. Une très large étude pharmacogénétique internationale (de pratiquement 2000 patients), a été réalisée en 2008 (Mallal et al., 2008). Elle a montré qu'en identifiant et excluant les patients porteurs de l'allèle HLA-B\*5701, la totalité des syndromes d'hypersensibilité liés au traitement disparaissait. La recherche de l'allèle HLA-B\*5701 est aujourd'hui requise par l'EMA et fortement recommandée par la FDA avant la mise sous abacavir des patients. Il a été démontré l'efficacité en matière de coûts de ce test génétique en pratique clinique (Kauf et al., 2010) et une augmentation de la prescription d'abacavir a été observée avec sa mise en place (Ingelman-Sundberg, 2008). Ceci montre que l'obligation de tests génétiques ne doit pas être vue comme un point négatif, notamment par l'industrie pharmaceutique.



**Figure 1.23.** Illustration de la relation génotype-phénotype (Eichelbaum et al., 2006).

- En haut les différents allèles du gène TPMT et en bas la distribution de l'activité de TPMT dans la population en fonction des allèles portés (wt : allèle sauvage, m : allèle muté). Les mutations sont responsables d'une diminution de l'activité de l'enzyme TPMT.
- Taux de nucléotides thioguanine (TGN) et fréquences de survenue de toxicité en fonction du génotype lorsque celui-ci n'est pas pris en compte dans le choix de la dose de thiopurine administrée aux patients.
- Taux de nucléotides thioguanine (TGN) et fréquences de survenue de toxicité en fonction du génotype lorsque la dose la dose de thiopurine administrée aux patients est individualisée selon leur génotype.

Un autre exemple illustrant l'association entre une variation génétique et la probabilité de survenue de toxicités concerne le gène TPMT codant pour la thiopurine S-méthyltransférase, une enzyme responsable du métabolisme des médicaments de la famille des thiopurines (azathiopine, mercaptopurine, thioguanine). Cet exemple permet d'illustrer l'intérêt de l'individualisation des traitements selon le génotype (**Figure 1.23**). Différents polymorphismes sont responsables d'une diminution de l'activité de l'enzyme TPMT. Ainsi lorsqu'une même dose de thiopurine est administrée aux patients, indifféremment du génotype pour le gène TPMT, les hétérozygotes présentent des taux de nucléotides thioguanine (TGN) deux fois plus élevés que les homozygotes sauvages et des taux 10 fois plus élevés sont retrouvés chez les homozygotes mutés. Ce qui se traduit par des fréquences de toxicité significativement plus élevées chez les porteurs d'un ou deux allèles mutés. À l'inverse, lorsque la dose est adaptée en fonction du génotype de TPMT des concentrations similaires de TGN sont retrouvées dans les différentes sous-populations génétiques. La probabilité de survenue de toxicité chez les hétérozygotes et les homozygotes mutés se ramène alors à celle des homozygotes sauvages (Eichelbaum et al., 2006).

Dans ces premiers exemples le polymorphisme était associé à un effet indésirable. De nombreux liens PGt ont également été révélés sur l'efficacité des traitements, notamment en oncologie. L'EMA a rendu obligatoire la détection d'une mutation du gène activant les récepteurs à l'EGF (*Epidermal Growth Factor Receptor, EGFR*) chez les patients présentant un stade avancé de cancer du poumon non à petites cellules avant la mise sous géfitinib. Cette mutation est en effet liée à une augmentation de la médiane de survie de 10 à 27 mois chez ces patients (Rosell et al., 2009). Tandis que d'autres mutations, notamment la mutation rs121434569 (mutation T790M pour la protéine), sont associées à une résistance au géfitinib (Yoshida et al., 2010) et une alternative thérapeutique doit être trouvée pour ces patients. Le test pharmacogénétique permet ici d'identifier les patients susceptibles de mieux répondre aux traitements. Toujours dans le domaine de l'oncologie, le tamoxifène est un traitement donné aux femmes atteintes d'un cancer du sein hormonodépendant. Plusieurs polymorphismes responsables d'une diminution ou d'une perte de fonction du CYP2D6 ont été associés à une diminution de la réponse (Schroth et al., 2009). Le CYP2D6 est responsable de la transformation du tamoxifène, une prodrogue inactive, en son métabolite actif, l'endoxifène, ayant une affinité 100 fois plus grande pour la cible, le récepteur aux œstrogènes (*Estrogens Receptor, ER*). Cet exemple met en évidence le lien PK/PD et montre

l'importance de l'impact de la génétique sur cette relation. Dans une autre aire thérapeutique, un lien a été mis en évidence entre un polymorphisme du gène *SLCO1B1* codant pour le transporteur *OATP1B1* et les myopathies induites par simvastatine, un traitement hypocholestérolémiant de la famille des statines. Le transporteur *OATP1B1* est notamment responsable de l'entrée des statines dans les cellules hépatiques (Kalliokoski and Niemi, 2009). Cette association a été mise en évidence dans une étude GWAS où 316184 marqueurs ont été étudiés chez 85 sujets présentant des myopathies et 90 sujets contrôles prenant de la simvastatine, et confirmée dans une étude de réplication incluant 16664 patients (SEARCH Collaborative Group et al., 2008). L'association a été faite avec un marqueur observé mais non causal (rs4363657) en fort déséquilibre de liaison avec le marqueur causal rs4149056, tous les deux positionnés sur le gène *SLCO1B1*. Bien que plus de 60% des myopathies induites par simvastatine soit attribuées à ce polymorphisme, la faible incidence de cet effet indésirable (1.5 à 5% (Joy and Hegele, 2009)) rend un test génétique de routine inefficace en matière de coûts.

Dans le cas d'associations mises en évidence sur des données PK, l'exemple le plus connu est celui de la warfarine, dont la dose à l'initiation du traitement est dépendante de plusieurs facteurs génétiques. Deux allèles sont responsables de la perte de fonction du *CYP2C9*, responsable du métabolisme de la warfarine, et sont donc à l'origine de surdosage (Aithal et al., 1999). Un autre polymorphisme, cette fois-ci lié à l'effet du traitement, a été mis en évidence sur le gène *VKORC1* codant pour la cible de la warfarine (Rieder et al., 2005). Différents algorithmes ont été développés pour déterminer la dose initiale optimale à administrer aux patients, basés sur des facteurs cliniques et génétiques (International Warfarin Pharmacogenetics Consortium et al., 2009). Il est intéressant de voir que ces algorithmes ne sont pas toujours applicables lorsque la population change, du fait que les fréquences des polymorphismes varient en fonction de l'origine ethnique. Ainsi un algorithme développé sur une population caucasienne britannique (Sconce et al., 2005) montrait de faibles performances lorsqu'appliqué à la population brésilienne (Botton et al., 2011; Suarez-Kurtz, 2011), connu pour sa très grande diversité ethnique et génétique (Suarez-Kurtz, 2010). Un autre exemple est celui de l'éfavirenz, une molécule antirétrovirale utilisée pour le traitement de l'infection par VIH, pour laquelle des polymorphismes du *CYP2B6* ont été associés à une diminution de sa clairance (Bertrand et al., 2014). Ainsi, les patients porteurs de ces polymorphismes ont en moyenne des concentrations plus élevées

de médicament et un risque potentiellement plus élevé de toxicités. À ce jour aucune identification de ces polymorphismes n'est demandée avant l'instauration du traitement (Résumés des caractéristiques du produit, ANSM, mise à jour le 10/01/2014). De même, cinq polymorphismes, tous marqueurs du gène *SLCO1B1*, codant le transporteur *OATP1B1*, ont été associés à une variation de la clairance du méthotrexate dans le traitement des leucémies aiguës lymphoblastiques chez l'enfant (Treviño et al., 2009). Les variations du gène *SLCO1B1* permettaient d'expliquer 9.3 à 11.3% de la variabilité interindividuelle de la clairance selon la cohorte étudiée. Cette association confirmée dans une autre étude (Lopez-Lopez et al., 2011) n'a, à ce jour, pas abouti à des recommandations d'usage. Enfin, toujours dans l'étude des associations entre les polymorphismes du gène *SLCO1B1* et la clairance du méthotrexate, une étude a montré que ce sont les variants les moins fréquents de *SLCO1B1* dans la population étudiée qui expliquaient la plus grande partie de la variabilité d'origine génétique du phénotype (Ramsey et al., 2012).

### **1.2.5. La pharmacogénétique dans le développement du médicament**

#### ***1.2.5.1. Le développement clinique d'un candidat médicament***

Le développement clinique d'un médicament se décompose en quatre phases. Sur les milliers de molécules identifiées à l'étape de recherche, seule une dizaine sont encore en développement à cette étape. Elles sont sélectionnées sur des modèles *in vitro*, *ex vivo* et *in vivo* (chez l'animal lors du développement préclinique). Parmi les dix candidats médicaments, un seul passera avec succès par toutes les phases du développement clinique jusqu'à l'obtention de l'autorisation de mise sur le marché (AMM).

Les phases du développement cliniques sont différentes en matière d'objectifs ou de population étudiée. De plus, pour l'étude de la pharmacocinétique chez l'Homme, la quantité de données disponible varie d'une phase à l'autre. La phase I est la première administration du candidat médicament chez l'Homme et se fait généralement chez des volontaires sains (sauf exception, *e.g.* en oncologie) afin d'évaluer la tolérance du traitement. Un nombre limité de sujets est inclus (moins de 100 sujets) avec des critères d'inclusion homogènes, afin de limiter la variabilité interindividuelle. Les doses de médicament sont augmentées progressivement en administration unique puis en administrations répétées. L'étude de la pharmacocinétique est un autre objectif de ces études de phase I où de nombreux prélèvements sont généralement effectués. La pharmacocinétique est alors analysée par NCA ou par régression non linéaire et la variabilité

des processus ADME est estimée par des approches en deux étapes ou au travers de modèles mixtes. C'est généralement au cours des études de phase II que le médicament est administré pour la première fois chez le malade, afin de montrer son efficacité thérapeutique (preuve de concept). Ce sont des essais de tailles réduites (100 à 500 patients) afin de conserver une homogénéité de la réponse. Les données récoltées à cette étape permettent d'évaluer la relation dose - concentrations - effets (PK/PD) pour notamment déterminer la dose optimale (étude de recherche de dose (Food and Drug Administration (FDA), 2003)), *i.e.* la dose associée à la plus grande efficacité thérapeutique avec le moins d'effets indésirables. Il y a donc encore, lors des essais de phase II, des prélèvements afin d'étudier la pharmacocinétique du médicament, mais le nombre de prélèvements est généralement limité. À cette étape les analyses PK individuelles et par extension les approches en deux étapes, qui requièrent de nombreux prélèvements par sujet, ne sont plus utilisables, à l'inverse de l'analyse par modèles mixtes. Les études de phase III servent à comparer le médicament en développement à un traitement de référence ou un placebo, et à confirmer la dose choisie pour le médicament. Ces essais tendent à se rapprocher des conditions réelles et incluent donc un très grand nombre de patients (500 à 5000), pour étudier l'efficacité du produit dans des conditions plus hétérogènes. Ces essais ne sont pas conçus pour l'étude de la pharmacocinétique et généralement très peu de données PK sont disponibles. Néanmoins les données PK en phase III sont très intéressantes pour étudier notamment l'effet des covariables, du fait du grand nombre de sujets. Après obtention de l'AMM et commercialisation du médicament, de nouveaux essais pourront être mis en place pour notamment évaluer une nouvelle formulation ou une nouvelle indication thérapeutique (phase IIIb).

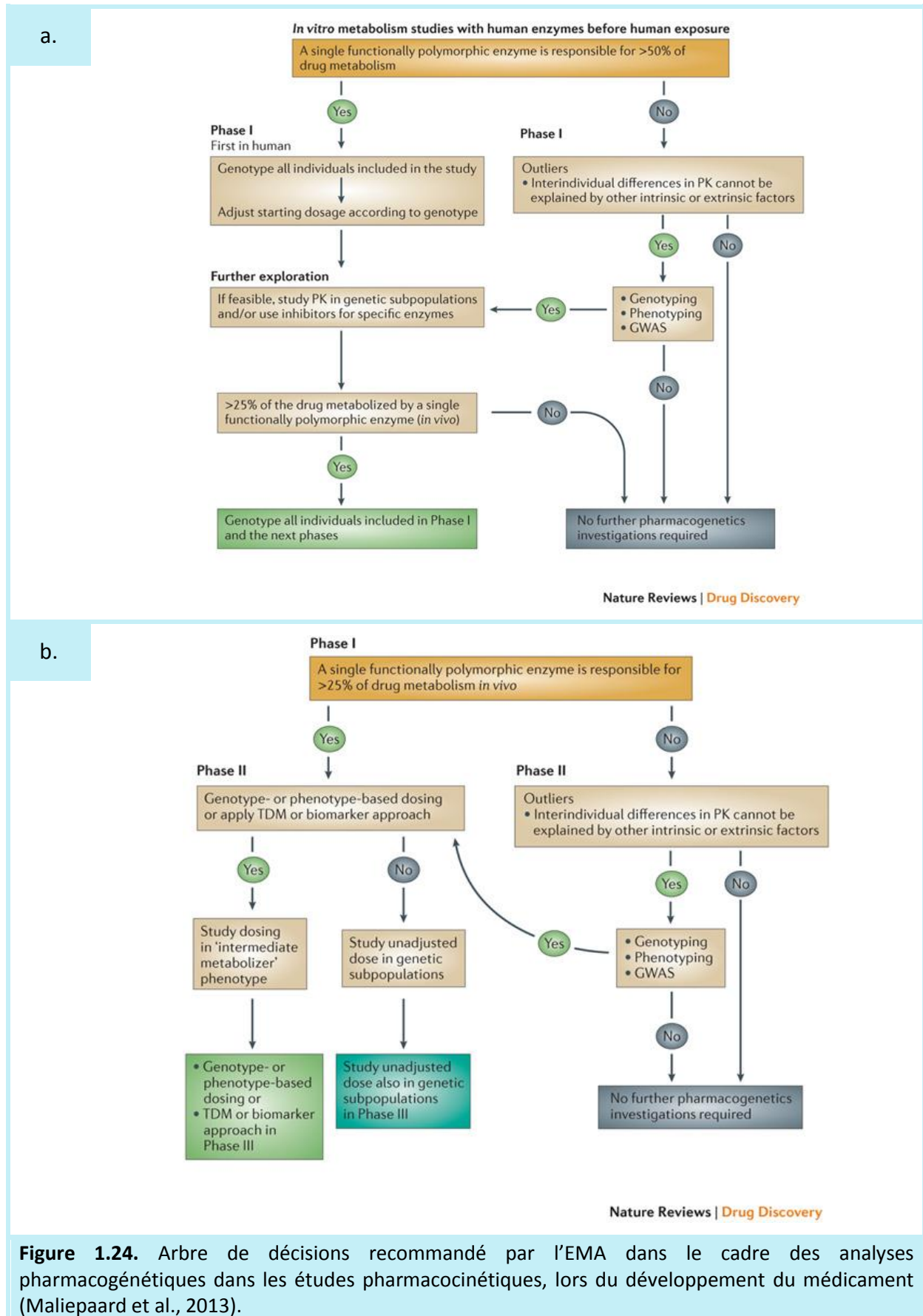
### ***1.2.5.2. Recommandations EMA/FDA pour l'intégration des analyses pharmacogénétiques dans le développement clinique des médicaments***

L'EMA et la FDA ont publié des recommandations sur l'utilisation de la pharmacogénétique dans le développement de nouvelles entités chimiques, et notamment lors des études pharmacocinétiques. L'EMA donnait en 2012 (European Medicines Agency (EMA), 2012) les situations où l'étude de la pharmacogénétique dans le développement clinique est requise ou recommandée, la distinction étant faite sur l'implication (en pourcentage) de l'enzyme ou transporteur polymorphique dans la pharmacocinétique de la nouvelle molécule. Pour résumé, ces situations sont les suivantes : lorsque les études *in vitro* ou *in vivo* précédentes

ont montré que des enzymes du métabolisme ou des transporteurs connus pour être polymorphiques représentent une voie importante de l'élimination ou de la distribution du médicament; ou lorsqu'une forte variabilité interindividuelle est observée et ne peut être expliquée par d'autres facteurs. L'EMA a également publié un arbre décisionnel sur la mise en place d'analyse PGt dans le développement clinique. Ces analyses doivent être faites au plus tôt dans le développement clinique, dès la phase I si une enzyme polymorphique est impliquée dans le métabolisme du médicament (**Figure 1.24a**). L'EMA recommande de continuer les génotypages dans les études suivantes, incluant les études de phase II et phase III, pour augmenter les données supportant de possibles recommandations d'usage du médicament. Si une association entre un ou plusieurs marqueurs génétiques et la variabilité PK est mise en évidence, un ajustement de la dose en fonction du génotype est recommandé pour les études suivantes (**Figure 1.24b**). Si l'intérêt d'étudier la pharmacogénétique d'une nouvelle molécule se révèle plus tardivement, par la mise en évidence d'une nouvelle enzyme ou transporteur polymorphique impliqué dans la pharmacocinétique du médicament ou par l'observation d'*outliers* (données individuelles largement séparées du groupe principal de données), l'étude de la pharmacogénétique est alors recommandée pour les essais suivants (**Figure 1.24a**). Dans ce cas l'utilisation d'une approche gènes candidats incluant un grand nombre de marqueurs génétiques est intéressante pour mettre en évidence une relation jusque-là non observée entre la génétique et la PK.

La FDA (*Food and Drug Administration*) a également émis des recommandations sur l'utilisation de la pharmacogénétique dans le développement clinique (Food and Drug Administration (FDA), 2013), centrées sur les phases précoces du développement (phase I et phase II). Les situations où l'étude de la pharmacogénétique est recommandée sont similaires à celles de l'EMA. La FDA met, de surcroît, en avant l'intérêt de l'étude précoce de la pharmacogénétique afin d'augmenter la probabilité de succès des essais cliniques ultérieurs.

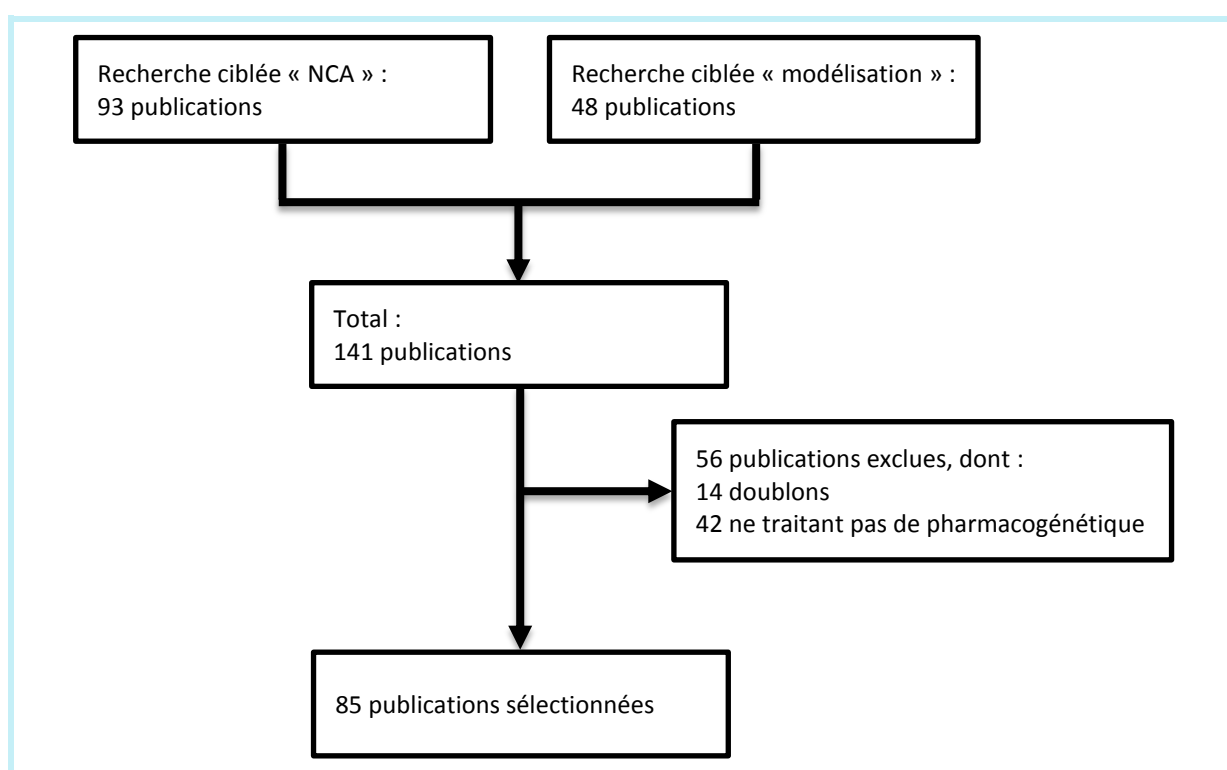




### 1.2.5.3. Données de la littérature sur les analyses pharmacogénétiques dans les études pharmacocinétiques

Une revue de la littérature a permis de faire le point sur la méthodologie utilisée pour les analyses PGt dans les études PK. Les résultats de cette revue ont été publiés en introduction des premiers travaux présentés dans le chapitre 4.

Cette revue faite à partir des publications rassemblées dans la base Pubmed était limitée aux articles publiés en anglais entre le 1<sup>er</sup> janvier 2010 et le 31 décembre 2012 sur des essais cliniques. 85 publications ont été sélectionnées à partir de deux chaînes de mots clés distinguant les analyses PGt basées sur des phénotypes estimés par NCA ou par modélisation (Figure 1.25).



**Figure 1.25.** Diagramme de sélection des publications pour la revue.

Recherche « NCA » : (((((polymorph\* OR genetic\* OR pharmacogenet\*[Title/Abstract])) AND (pharmacokinet\* AND concentration\*[Title/Abstract])) AND (non-compartment\* OR noncompartment\* OR NCA OR AUC\*[Title/Abstract])) AND Clinical Trial[ptyp] AND hasabstract[text] AND ( "2010/01/01"[PDat] : "2012/12/31"[PDat] ) AND Humans[Mesh] AND English[lang])

Recherche « modélisation » : (((((polymorph\* OR genetic\* OR pharmacogenet\*[Title/Abstract])) AND (pharmacokinet\* AND concentration\*[Title/Abstract])) AND ("population PK" OR "population pharmacokinetic" OR model\* OR nonlinear OR non-linear OR mixed effect\* OR Nonmem OR Monolix[Title/Abstract])) AND Clinical Trial[ptyp] AND hasabstract[text] AND ( "2010/01/01"[PDat] : "2012/12/31"[PDat] ) AND Humans[Mesh] AND English[lang])

Sur les 85 publications trouvées, plus des deux tiers basaient leurs analyses PGt sur un phénotype estimé par NCA (69%, généralement sur l'AUC), les 31% restant utilisant un phénotype estimé par modélisation. Un nombre limité de polymorphismes était étudié, en moyenne 3 gènes (avec une échelle de 1 à 45 gènes) et 8 SNPs (avec une échelle de 1 à 198 SNPs), montrant que les approches gènes candidats sont préférées car plus adaptées aux études PK. En matière de nombre de sujets, la majorité des analyses incluait un nombre limité de sujets (moins de 50 sujets, 66%) tandis que seules 15% des analyses incluaient plus de 100 sujets. Enfin différentes méthodes pour les analyses statistiques ont été utilisées : descriptives, univariées ou multivariées.

Cette revue montre donc l'hétérogénéité de la méthodologie des analyses PGt dans les études PK. Il n'y a pas de consensus mais la majorité des analyses utilise un phénotype estimé par NCA et sont effectuées sur un petit effectif, les analyses pharmacogénétiques se faisant généralement lors de la phase I du développement clinique des médicaments.

# Chapitre 2

## PRÉSENTATION DU CAS RÉEL

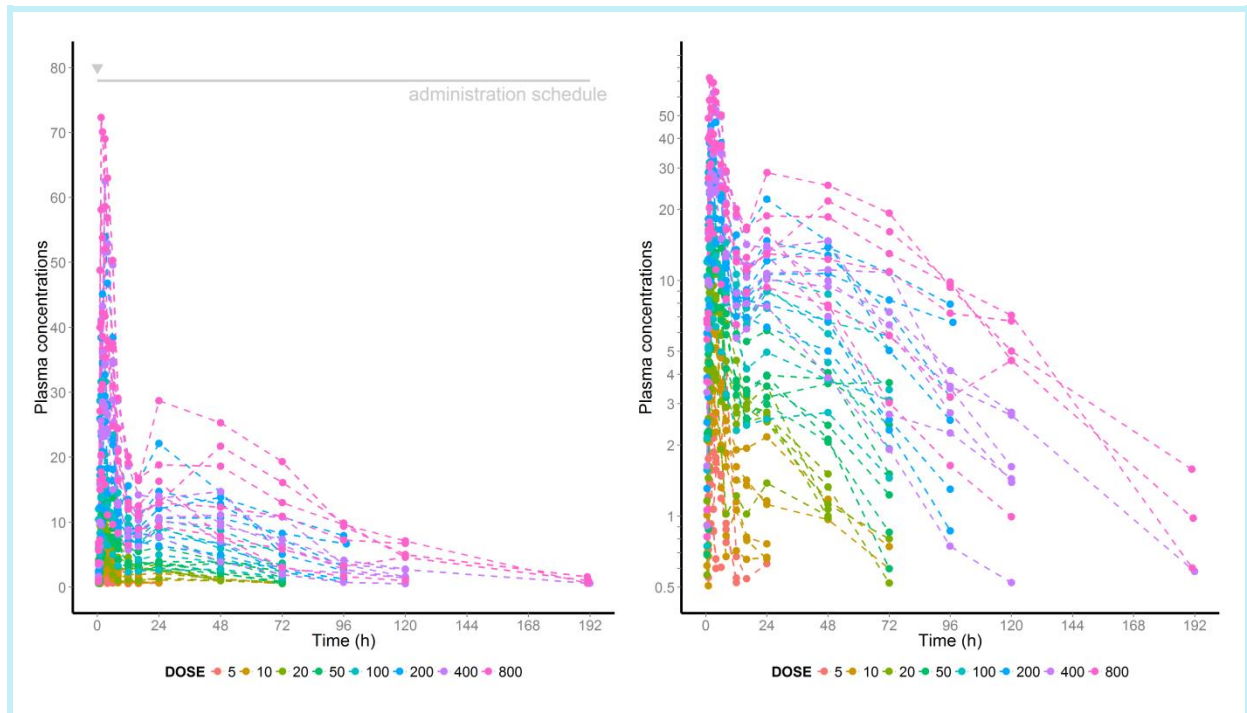
Cette thèse est issue d'une collaboration entre un laboratoire de recherche académique (IAME INSERM UMR1137/Université Paris Diderot) et l'industrie pharmaceutique (Département de Pharmacocinétique Clinique et de Pharmacométrie, Institut de Recherches Internationales Servier, IRIS). IRIS a notamment fourni un exemple de données réelles ayant motivé et servi de base aux différents travaux. Nous présentons dans ce chapitre cet exemple.

### 2.1. Données

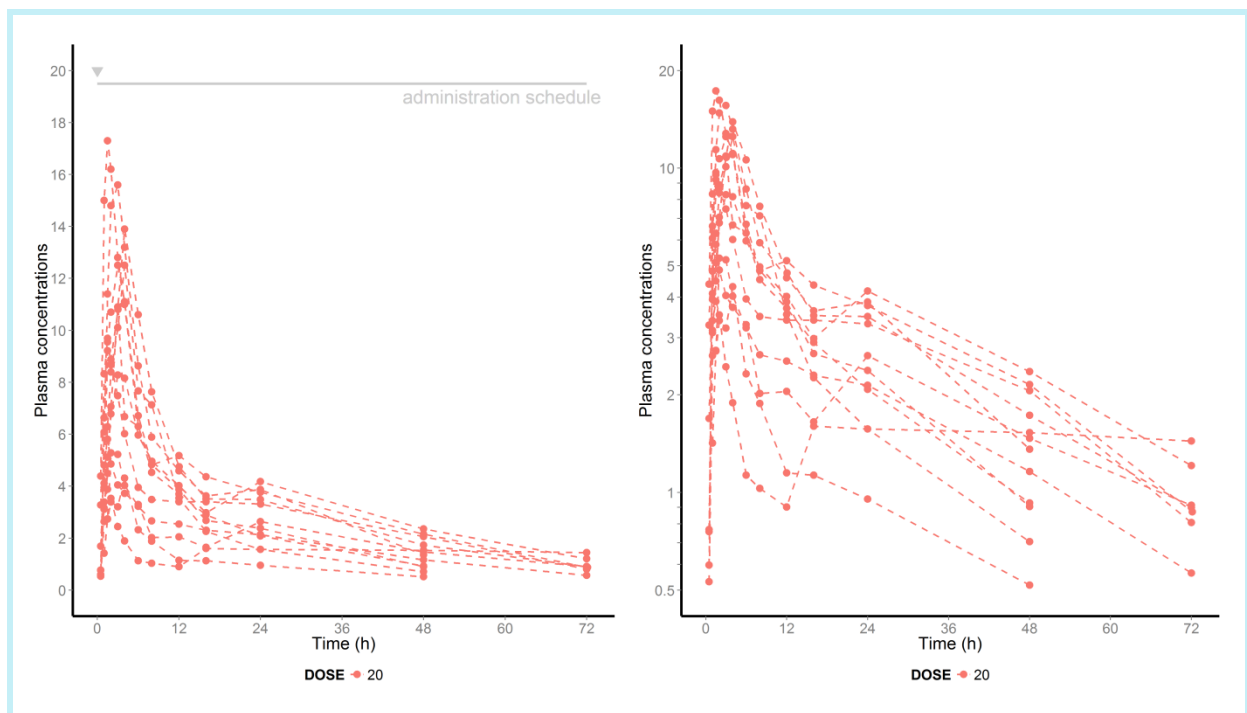
Une molécule que nous nommons « S » est développée par IRIS. Pour des raisons de confidentialité, la plupart des informations concernant cette molécule ne peuvent être renseignées. Les données de trois études de phase I, incluant au total 78 adultes volontaires sains, sont disponibles et servent de base aux travaux. Les caractéristiques de ces trois études sont les suivantes :

La première étude est une étude d'escalade de dose et inclue 48 sujets recevant la molécule S en dose unique (**Figure 2.1**). Au total 8 doses ont été administrées (5, 10, 20, 50, 100, 200, 400 et 800 unités) par voie orale sous la forme d'un comprimé, avec 6 sujets par dose. De nombreux prélèvements sanguins sont effectués pour mesurer les concentrations plasmatiques de la molécule S : avant l'administration puis 0,5, 1, 1,5, 2, 3, 4, 6, 8, 12, 16, 24, 48 et 72 heures après l'administration de la molécule S. Pour les plus fortes doses (200 unités et plus), des prélèvements supplémentaires sont faits 96, 120 et 192 heures après l'administration de la molécule S.

Dans la seconde étude, 12 sujets reçoivent une dose orale unique de la molécule S sous la forme d'un comprimé, à la dose de 20 unités (**Figure 2.2**). Des prélèvements sont effectués avant l'administration puis 0,5, 1, 1,5, 2, 3, 4, 6, 8, 12, 16, 24, 48 et 72 heures après l'administration de la molécule S.

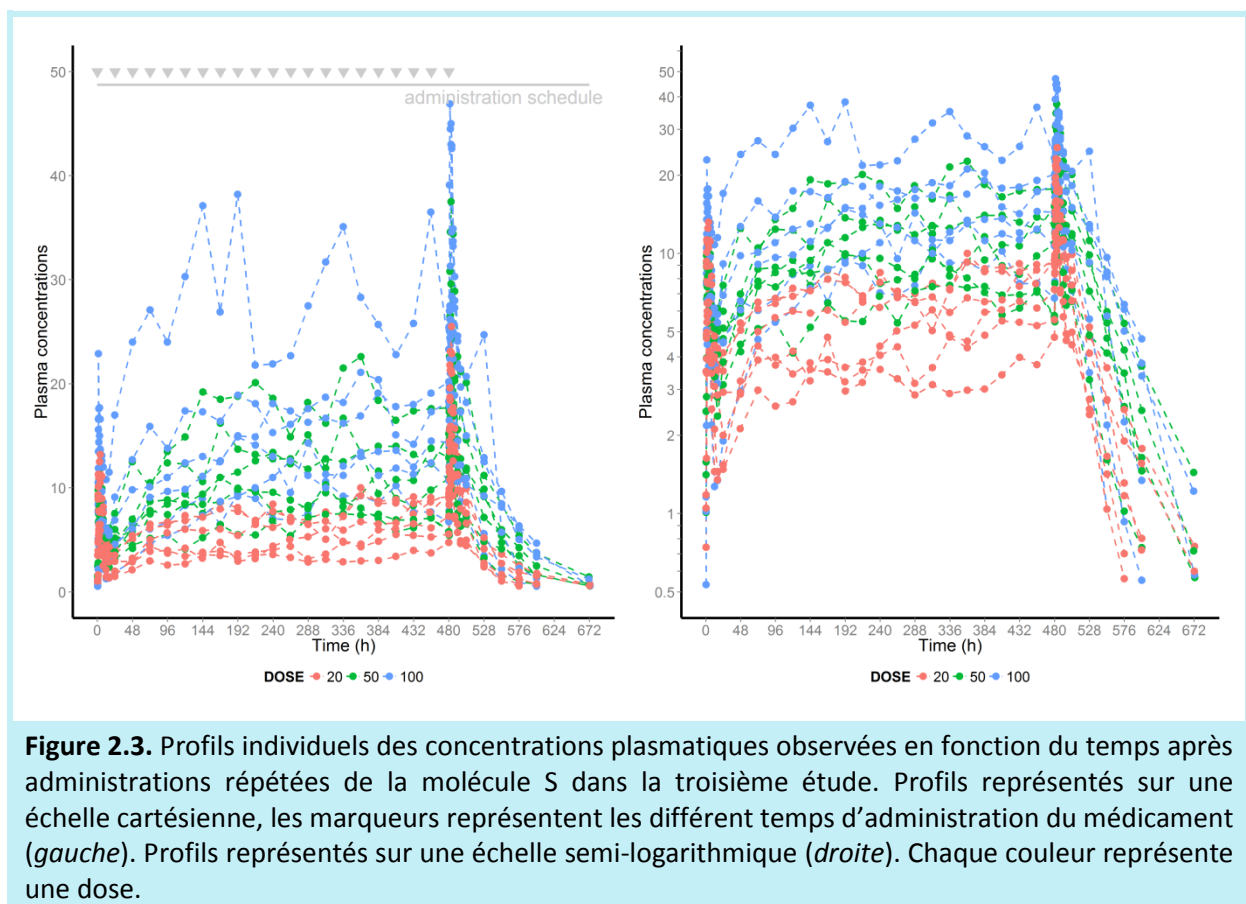


**Figure 2.1.** Profils individuels des concentrations plasmatiques observées en fonction du temps après administration unique de la molécule S dans la première étude. Profils représentés sur une échelle cartésienne, le marqueur représente le temps d'administration du médicament (*gauche*). Profils représentés sur une échelle semi-logarithmique (*droite*). Chaque couleur représente une dose.



**Figure 2.2.** Profils individuels des concentrations plasmatiques observées en fonction du temps après administration unique de la molécule S dans la deuxième étude. Profils représentés sur une échelle cartésienne, le marqueur représente le temps d'administration du médicament (*gauche*). Profils représentés sur une échelle semi-logarithmique (*droite*).

La troisième étude est une étude d'escalade de dose en doses répétées, où le médicament est administré par voie orale une fois par jour (**Figure 2.3**). Cette étude compte 18 sujets recevant 3 doses différentes (20, 50 et 100 unités), avec 6 sujets par dose. Plusieurs prélèvements sont effectués le premier jour, avant et 0,5, 1, 1,5, 2, 3, 4, 6, 8, 12 et 16 heures après la première prise du comprimé. Du jour 2 au jour 20, un prélèvement avant l'administration est effectué pour mesurer la concentration résiduelle. Enfin, des prélèvements sont faits avant l'administration, puis 0,5, 1, 1,5, 2, 3, 4, 6, 8, 12, 16, 24, 48, 72, 96, 120 et 192 heures après la dernière administration de la molécule S, au jour 21.



Tous les sujets de ces trois études ont été génotypés à l'inclusion grâce à une puce à ADN développée par IRIS, spécialement pour les études PK. En effet cette puce comprend des polymorphismes de gènes codant pour des protéines impliquées dans la PK des médicaments (**Table 2.1**), *i.e.* des enzymes du métabolisme, des transporteurs et des récepteurs nucléaires, présentés dans le paragraphe 1.1.3.

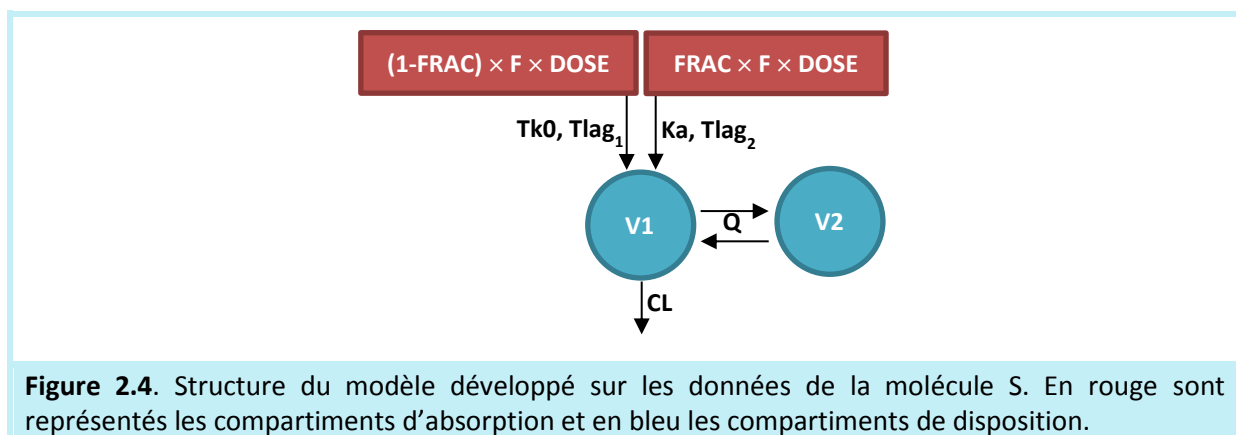
**Table 2.1.** Liste des gènes étudiés par la puce ADN développée par IRIS.  
Gènes (nombre de polymorphismes génotypés).

Enzymes de phase I	Enzymes de phase II	Transporteurs	Récepteurs nucléaires
CYP1A1 (7)	COMT (1)	SLC15A2 (NTCP, 4)	NR1I1 (PXR, 5)
CYP1A2 (3)	GST $\mu$ 1 (2)	SLC22A1 (OCT1, 11)	NR1I3 (CAR, 1)
CYP2A6 (8)	GSTp1 (2)	SLC22A2 (OCT2, 4)	NR1I1 (VDR, 3)
CYP2B6 (7)	GSTT1 (1)	SLC22A6 (OAT1, 1)	NFE2L2 (NRF2, 1)
CYP2C8 (6)	NAT1 (7)	SLCO1B1 (OATP1B1, 4)	NR2B1 (RXRA, 3)
CYP2C9 (12)	NAT2 (6)	SLCO1B3 (OATP1B3, 2)	
CYP2C19 (9)	SULT1A1 (4)	SLCO2B1 (OATP2B1, 1)	
CYP2D6 (19)	TPMT (5)	ABCB1 (P-GP, 42)	
CYP2E1 (1)	UGT1A1 (6)	ABCC2 (MRP2, 6)	
CYP3A4 (6)	UGT1A3 (1)	ABCG2 (BCRP, 4)	
CYP3A5 (5)	UGT1A6 (3)		
DPYD (5)	UGT1A7 (1)		
	UGT1A8 (2)		
	UGT1A9 (1)		
	UGT1A10 (2)		
	UGT2B7 (2)		
	UGT2B15 (1)		

Tableau modifié de documents internes à IRIS.

## 2.2. Analyse pharmacocinétique

Un modèle PK a été développé par approche de population sur les données des trois études présentées ci-dessus. La structure de ce modèle comprend deux compartiments de disposition et deux voies d'absorption pour décrire le rebond dans les concentrations observé sur les profils individuels. La première absorption suit un ordre 0 et la seconde un ordre 1. L'élimination est linéaire (**Figure 2.4**).



**Figure 2.4.** Structure du modèle développé sur les données de la molécule S. En rouge sont représentés les compartiments d'absorption et en bleu les compartiments de disposition.

La pharmacocinétique de la molécule S est non linéaire, ceci étant dû à sa faible solubilité. Pour prendre en compte cette non linéarité, les paramètres d'absorption FRAC et F, définis par la suite, sont non linéaires en fonction de la dose. La solution analytique caractérisant ce modèle est représenté dans la **Figure 2.5**.

$$f(\theta_i, Dose_i, t_{ij}) = C_0(t) + C_1(t),$$

$$C_0(t) = \begin{cases} 0 & \text{si } t \leq Tlag.Tk0, \\ \frac{F \times (1-FRAC) \times Dose}{Tk0} \left[ \frac{A0}{\alpha} (1-e^{-\alpha(t-Tlag.Tk0)}) + \frac{B0}{\beta} (1-e^{-\beta(t-Tlag.Tk0)}) \right] & \text{si } Tlag.Tk0 < t \leq Tk0, \\ \frac{F \times (1-FRAC) \times Dose}{Tk0} \left[ \frac{A0}{\alpha} (1-e^{-\alpha Tk0}) e^{-\alpha(t-Tlag.Tk0-Tk0)} + \frac{B0}{\beta} (1-e^{-\beta Tk0}) e^{-\beta(t-Tlag.Tk0-Tk0)} \right] & \text{sinon.} \end{cases}$$

$$C_1(t) = \begin{cases} 0 & \text{si } t \leq Tlag.Ka, \\ F \times FRAC \times Dose \left[ \frac{A1 e^{-\alpha(t-Tlag.Ka)} + B1 e^{-\beta(t-Tlag.Ka)}}{-(A1+B1)e^{-Ka(t-Tlag.Ka)}} \right] & \text{sinon.} \end{cases}$$

$$\alpha = \frac{\frac{Q}{V2} \frac{CL}{V1}}{\beta}, \quad \beta = \frac{1}{2} \left[ \frac{Q}{V1} + \frac{Q}{V2} + \frac{CL}{V1} - \sqrt{\left( \frac{Q}{V1} + \frac{Q}{V2} + \frac{CL}{V1} \right)^2 - 4 \frac{Q}{V2} \frac{CL}{V1}} \right],$$

$$A0 = \frac{1}{V1} \frac{\alpha - \frac{Q}{V2}}{\alpha - \beta}, \quad B0 = \frac{1}{V1} \frac{\beta - \frac{Q}{V2}}{\beta - \alpha},$$

$$A1 = \frac{ka}{V1} \frac{\frac{Q}{V2} - \alpha}{(Ka - \alpha)(\beta - \alpha)}, \quad B1 = \frac{ka}{V1} \frac{\frac{Q}{V2} - \beta}{(Ka - \beta)(\alpha - \beta)}.$$

$$F = \begin{cases} 1 & \text{si } Dose \leq 20 \\ 1 - \frac{(Dose - 20) Imax_F}{Dose - 20 + D_{50F}} & \text{sinon,} \end{cases} \quad FRAC = \frac{Dose \times Emax_{FRAC}}{Dose + D_{50FRAC}}$$

**Figure 2.5.** Décomposition de la solution analytique caractérisant le modèle développé sur les données de la molécule S.

Les paramètres du modèle sont les suivants : F est la biodisponibilité dépendante de la dose (F est constante pour les doses inférieures à 20 unités et décroît pour les doses supérieures selon un modèle Imax avec les paramètres Imax<sub>F</sub> et D<sub>50F</sub>), FRAC la fraction de la dose se divisant dans chaque voie d'absorption dépendante de la dose (FRAC croît avec dose selon un modèle Emax avec les paramètres Emax<sub>FRAC</sub> et D<sub>50FRAC</sub>), Tk0 la durée de l'absorption d'ordre 0 (équivalente à une perfusion à débit constant) et Tlag<sub>1</sub> le temps de latence correspondant, Ka la constante d'absorption d'ordre 1 et Tlag<sub>2</sub> le temps de latence



correspondant, V1 le volume du compartiment central, V2 le volume du compartiment périphérique, Q la clairance intercompartimentale et CL la clairance d'élimination.

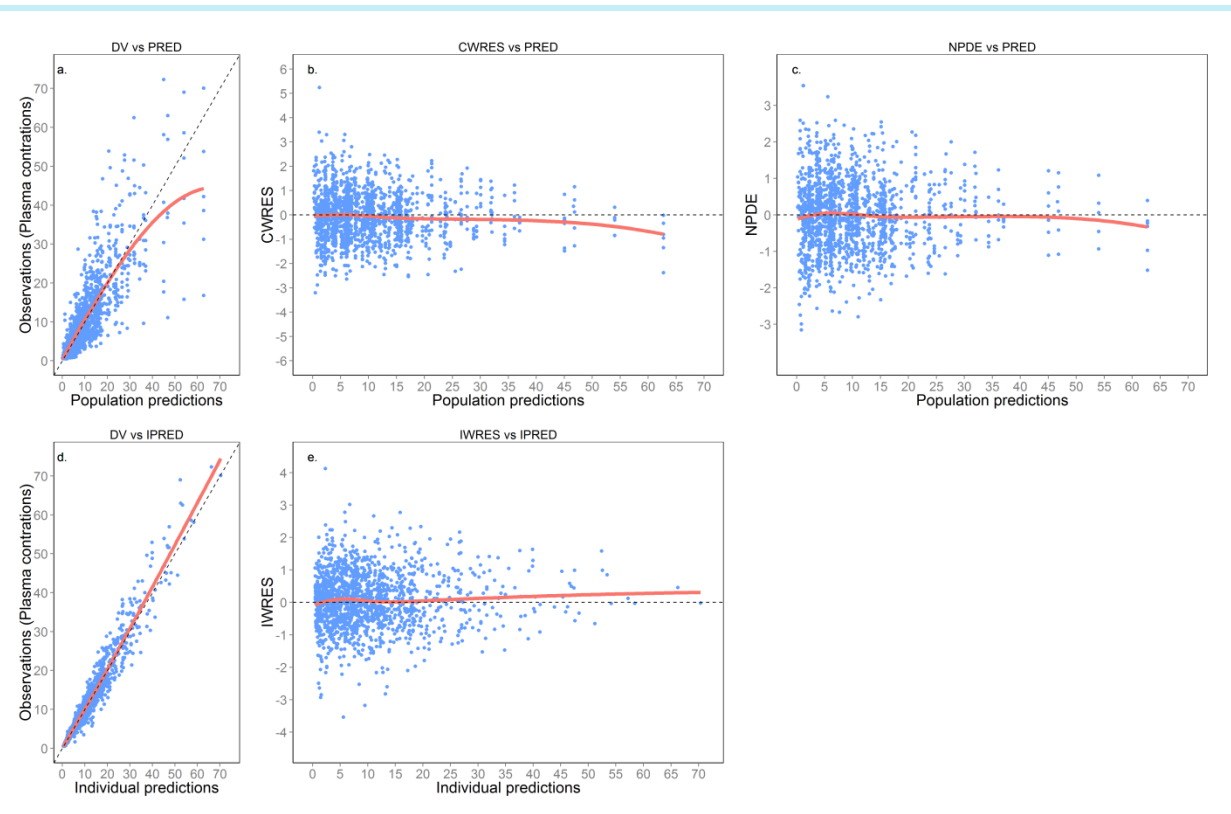
Les effets aléatoires sont modélisés avec un modèle exponentiel et l'erreur résiduelle est proportionnelle. Les estimations des paramètres de population du modèle final, et leurs erreurs standards, sont présentées dans la **Table 2.2**. La variabilité interindividuelle est modérée pour la plupart des paramètres (de 25.1% pour CL à 44.2% pour V2), avec une variabilité plus importante pour la clairance intercompartimentale (Q, 89.9%). L'erreur résiduelle proportionnelle est également modérée (20%). Tous les effets fixes sont estimés avec une bonne précision (RSE inférieures à 20%). Les estimations des variances des effets aléatoires sont associées à des RSE inférieures à 40%, hormis pour celle du paramètre Tk0 (RSE de 55%).

**Table 2.2.** Estimations des paramètres du modèle développé sur les données de la molécule S.

Paramètre		Effet fixe $\mu$ (SE)		Variance des effets aléatoires $\omega^2$ (SE)		Variabilité interindividuelle $\omega$ en %
F <sup>a</sup>	Imax <sub>F</sub>	0.800	(0.0317)	0.108	(0.0390)	32.9
	D <sub>50F</sub>	41.7	(4.72)			
FRAC <sup>b</sup>	E <sub>max</sub> <sub>FRAC</sub>	0.450	(0.0177)			
	D <sub>50FRAC</sub>	18.6	(3.36)			
Tlag <sub>1</sub>		0.401	(0.0226)	0.123	(0.0274)	35.1
Tk0		1.59	(0.0226)	0.100	(0.0552)	31.6
Tlag <sub>2</sub>		22.7	(0.288)			
Ka		0.203	(0.0264)			
V1		1520	(61.6)			
Q		147	(17.6)	0.808	(0.273)	89.9
V2		2130	(29.0)	0.195	(0.0656)	44.2
CL		94.9	(2.82)	0.0630	(0.0206)	25.1
$\sigma_{\text{slope}}$ (%)		20.0	(0.357)			
a. Pour des doses < 20 unités : $F = 1$ , pour des doses $\geq 20$ unités : $F = 1 - \frac{Imax_F (dose - 20)}{D50_F + dose - 20}$						
b. $FRAC = \frac{Emax_{FRAC} dose}{D50_{FRAC} + dose}$						
SE : Erreur standard						

Différents graphiques diagnostics sont représentés dans la **Figure 2.6**. Les graphiques comparant les concentrations observées (DV) aux prédictions de population (PRED) ou aux prédictions individuelles du modèle (IPRED) indiquent que ces prédictions sont en adéquation avec les observations. Les graphiques des résidus de population (CWRES) et des

erreurs de prédiction de distribution normalisées (NPDE), issues de 1000 simulations, en fonction des prédictions de population (PRED) et des résidus individuels (IWRES) en fonction des prédictions individuelles (IPRED), montrent des résidus ou NPDE distribués de façon homogène autour de l'axe à l'ordonnée égale à zéro. Ce qui montre également une bonne description des données par le modèle.



**Figure 2.6.** Graphiques diagnostics du modèle final.

- a. Concentrations observées (DV) en fonction des prédictions de population (PRED).
- b. Erreurs de prédiction de distribution normalisées (NPDE) en fonction des prédictions de population (PRED).
- c. Résidus conditionnels de population pondérés (CWRES) en fonction des prédictions de population (PRED).
- d. Concentrations observées (DV) en fonction des prédictions individuelles (IPRED).
- e. Résidus individuels pondérés (IWRES) en fonction des prédictions individuelles (IPRED).

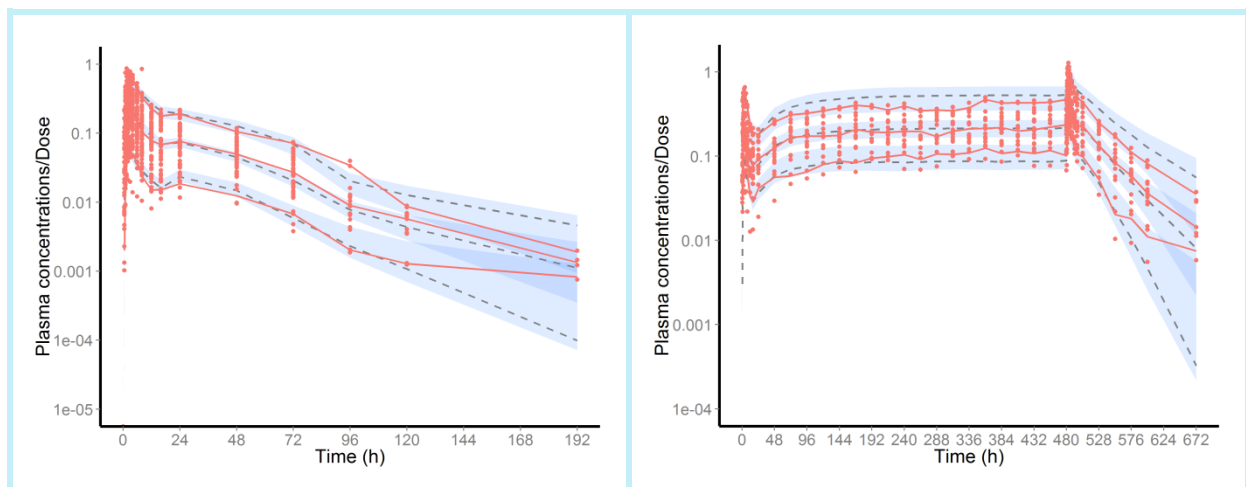
Les données sont représentées dans tous les graphiques par des points bleus.

La droite d'identité (graphiques a, d) et l'abscisse (graphiques b, c, e) sont représentées par une droite pointillée.

Un profil lissé des données est représenté dans tous les graphiques par un trait plein rouge.

Par ailleurs, 1000 jeux de données ont été simulés avec le modèle final et le protocole des études. La comparaison entre ces simulations et les données observées est effectuée à travers un VPC présenté dans la **Figure 2.7**. Afin de représenter l'ensemble des données, les

concentrations observées ou simulées sont normalisées par la dose. Le modèle décrit correctement les concentrations observées après administration unique, la médiane et les percentiles des observations étant proches de ceux obtenus par simulations. L'augmentation importante de l'intervalle de prédiction autour des percentiles des données simulées pour les temps tardifs est due au fait que peu de données sont disponibles à ces temps. Pour les concentrations observées après administrations répétées, le modèle sur-prédit le 95<sup>ème</sup> percentile des observations, tandis que la médiane et le 5<sup>ème</sup> percentile sont correctement prédits. Ceci peut être lié à une légère sur-estimation de la variabilité interindividuelle.



**Figure 2.7.** Visual predictive check (VPC) des concentrations plasmatiques de la molécule S, normalisées par la dose, des données après administration unique (*gauche*) et administrations répétées (*droite*), pour le modèle final. En rouge sont représentés : les concentrations observées, représentées par des points, et les 5<sup>ème</sup>, 50<sup>ème</sup> et 95<sup>ème</sup> percentiles de ces observations, représentés par des traits pleins. En gris les 5<sup>ème</sup>, 50<sup>ème</sup> et 95<sup>ème</sup> percentiles des concentrations simulées sont représentés par des traits pointillés et leurs intervalles de prédiction à 95% respectifs sont représentés par les aires bleues.

Finalement, les différentes évaluations graphiques indiquent que le modèle décrit correctement les concentrations observées de la molécule S et ce pour les différentes doses et schémas d'administration. Le protocole des études, la conception de la puce à ADN, ainsi que le modèle PK développé pour la molécule S ont été utilisés comme base pour nos travaux de simulations.

# Chapitre 3

## OBJECTIFS DE LA THÈSE

La revue de la littérature présentée au paragraphe 1.2.5.3 a montré que la majorité des analyses pharmacogénétiques dans les études PK se font avec des phénotypes estimés par NCA, malgré l'utilisation croissante des méthodes basées sur les MNLEM, et ce sur un petit nombre de sujets. Nous avons donc cherché à évaluer cette méthodologie par des simulations, afin de fournir des recommandations pour améliorer la détection des variants génétiques le cas échéant. Dans ces travaux, la méthodologie des analyses PGt a été évaluée sous trois aspects : les méthodes d'estimation du phénotype PK, les tests statistiques d'association et le protocole des études comprenant le nombre de sujets et le nombre de prélèvements. Le cas réel présenté précédemment fournit un contexte intéressant de développement clinique où de nombreux polymorphismes ont été génotypés avec une approche gènes candidats, ciblée sur la pharmacocinétique des médicaments.

Dans les premiers travaux, présentés au chapitre 4, nous comparons la capacité de différents phénotypes PK, observés ou estimés par NCA ou MNLEM, à capturer l'information nécessaire pour détecter l'impact de plusieurs variants génétiques. Nous évaluons également la performance de différents tests d'association (une *stepwise procedure* et trois régressions pénalisées : la régression *ridge*, le Lasso et l'HyperLasso) pour détecter ces polymorphismes. Ces simulations s'inscrivent dans un contexte de phase I et l'effet d'un faible effectif, typique de ces études, sur la puissance de l'essai pour détecter l'impact de la génétique est également évalué.

Les seconds travaux présentés au chapitre 5 s'inscrivent dans la continuité des premiers. Afin d'améliorer l'identification des variants génétiques, nous évaluons l'effet de combiner des données d'études de phase I et de phase II sur la puissance de détection des polymorphismes génétiques. L'intérêt de l'optimisation des protocoles de phase II, épars, est également évalué.

Afin d'augmenter la puissance de détection des polymorphismes dans les études PK, nous évaluons l'intérêt de prendre en compte l'intégralité de la distribution conditionnelle des

paramètres individuels dans les MNLEM, dans les tests d'association. Cette étude de simulation est présentée dans le chapitre 6.

# Chapitre 4

## COMPARAISON DES MÉTHODES D'ESTIMATION DES PHÉNOTYPES PHARMACOCINÉTIQUES ET DE TESTS D'ASSOCIATION POUR LES ANALYSES PHARMACOGÉNÉTIQUES

### 4.1. Résumé

La revue de la littérature présentée au paragraphe 1.2.5.3 montre que la plupart des analyses PGt dans les études PK publiées entre 2010 et 2012 utilisent un phénotype estimé par NCA comme l'AUC (plus des deux tiers). De plus, environ deux-tiers de ces analyses incluent moins de 50 sujets et seules 15% incluent plus de 100 sujets. Cette revue montre donc une absence de consensus quant à la méthodologie à utiliser pour les analyses PGt.

Différents tests d'association peuvent être utilisés pour mettre en évidence l'effet d'un variant génétique sur un phénotype PK. Les approches *stepwise procedure* sont usuelles dans le développement des modèles de covariables dans les analyses basées sur les MNLEM. Elles peuvent être opposées à des approches en une étape, plus récentes et complexes, que sont les régressions pénalisées.

Nous proposons donc dans ce premier travail de simulation de comparer à une *stepwise procedure* basée sur de simples régressions univariées (inspirée de (Lehr et al., 2010)), trois approches basées sur des régressions pénalisées (la régression *ridge* (Cule and De Iorio, 2013), le Lasso (Tibshirani, 1994) et l'HyperLasso (Hoggart et al., 2008)). Nous nous intéressons aux performances de ces tests d'association pour détecter plusieurs variants génétiques influençant la pharmacocinétique d'une molécule. Quatre phénotypes sont également évalués, deux concentrations observées et deux phénotypes estimés par NCA ou MNLEM, respectivement l'AUC et les estimations individuelles de la clairance.

Les simulations s'inspirent du cas réel présenté au chapitre 2, *i.e.* une molécule en développement clinique présentant une pharmacocinétique non linéaire (absorption

dépendante de la dose) et pour laquelle les données PK de trois études de phase I, ainsi que les données de 176 polymorphismes (SNPs), sont disponibles. Nous simulons donc, à l'aide du modèle développé pour la molécule, des profils PK dans deux scénarios incluant respectivement 78 sujets, comme pour les études de phase I, et 384 sujets, pour se rapprocher des conditions asymptotiques. Les génotypes de 176 SNPs sont également simulés à l'aide d'une base de référence Hapmap (International HapMap Consortium, 2003), afin de conserver les corrélations entre les différents polymorphismes. Sous l'hypothèse nulle  $H_0$ , la PK est simulée sans effet de la génétique alors que sous l'hypothèse alternative  $H_1$ , six SNPs sont tirés aléatoirement pour affecter la clairance avec un effet additif. L'effet de chaque SNP causal est calculé en fonction de la fréquence du polymorphisme et de la part de la variance de la clairance expliquée par le variant (Bertrand and Balding, 2013). Les six variants génétiques causaux expliquent respectivement 1, 2, 3, 5, 7 et 12% de la variabilité interindividuelle de la clairance. À partir des profils PK simulés sous  $H_0$  ou  $H_1$  sont alors déterminés les quatre phénotypes. Finalement les associations entre ces différents phénotypes et les 176 polymorphismes sont testées par *stepwise procedure* ou une des trois régressions pénalisées. Pour chacune des seize combinaisons entre les quatre phénotypes et les quatre tests, l'erreur de type I globale (FWER) sous  $H_0$  et la probabilité de détecter les effets génétiques sous  $H_1$  sont évaluées.

Dans les différents tests d'association, le FWER ciblé est de 20% et l'erreur de type I par test est calibrée par une correction de Šidák. Sous  $H_0$ , le FWER estimé s'avère significativement inférieur aux 20% attendus et une nouvelle correction empirique est utilisée pour en assurer son contrôle avant le calcul de puissance sous  $H_1$ . Une nouvelle simulation est effectuée avec des polymorphismes non corrélés pour étudier spécifiquement ce point. Dans ce cas, le FWER est correctement contrôlé, suggérant que la corrélation entre les variants génétiques est à l'origine de la diminution de l'erreur de type I globale.

Sous  $H_1$ , la comparaison de la puissance pour détecter les variants causaux en fonction du phénotype, quantifiée par le taux de vrais positifs, montre que dans les deux scénarios et pour les quatre tests d'association, le phénotype estimé par MNLEM permet une plus grande puissance que les phénotypes observés ou estimés par NCA. Des profils PK sont alors simulés avec un modèle simplifié n'incluant pas de non linéarité et de variabilité sur l'absorption. Dans ce cas, le nombre de vrais positifs est égal pour l'AUC et la clairance, respectivement

estimée par NCA et par MNLEM. La probabilité de détecter les différents variants causaux est similaire entre les différents tests d'association, cependant la régression *ridge* a tendance à détecter un nombre plus important de vrais positifs mais également de faux positifs. Cette puissance est faible dans le scénario de phase I incluant un nombre limité de sujets (78), tandis que l'augmentation importante du nombre de sujets dans le second scénario (384) entraîne une amélioration significative de la puissance de détection et représente le seul moyen de détecter plusieurs des six variants causaux.

Ce travail fait l'objet d'une publication dans *The AAPS Journal*.

#### **4.2. Article 1 (publié)**



Research Article

## Comparison of Nonlinear Mixed Effects Models and Noncompartmental Approaches in Detecting Pharmacogenetic Covariates

Adrien Tessier,<sup>1,2,3,6</sup> Julie Bertrand,<sup>4</sup> Marylore Chenel,<sup>3</sup> and Emmanuelle Comets<sup>1,2,5</sup>

Received 1 August 2014; accepted 28 January 2015; published online 20 February 2015

**Abstract.** Genetic data is now collected in many clinical trials, especially in population pharmacokinetic studies. There is no consensus on methods to test the association between pharmacokinetics and genetic covariates. We performed a simulation study inspired by real clinical trials, using the pharmacokinetics (PK) of a compound under development having a nonlinear bioavailability along with genotypes for 176 single nucleotide polymorphisms (SNPs). Scenarios included 78 subjects extensively sampled (16 observations per subject) to simulate a phase I study, or 384 subjects with the same rich design. Under the alternative hypothesis ( $H_1$ ), six SNPs were drawn randomly to affect the log-clearance under an additive linear model. For each scenario, 200 PK data sets were simulated under the null hypothesis (no gene effect) and  $H_1$ . We compared 16 combinations of four association tests, a stepwise procedure and three penalised regressions (ridge regression, Lasso, HyperLasso), applied to four pharmacokinetic phenotypes, two observed concentrations, area under the curve estimated by noncompartmental analysis and model-based clearance. The different combinations were compared in terms of true and false positives and probability to detect the genetic effects. In presence of nonlinearity and/or variability in bioavailability, model-based phenotype allowed a higher probability to detect the SNPs than other phenotypes. In a realistic setting with a limited number of subjects, all methods showed a low ability to detect genetic effects. Ridge regression had the best probability to detect SNPs, but also a higher number of false positives. No association test showed a much higher power than the others.

**KEY WORDS:** noncompartmental analysis; nonlinear mixed effects models; penalised regression; pharmacogenetics; pharmacokinetics.

### INTRODUCTION

Personalized care development should improve efficacy of drugs, and limit the risks associated with their use (1), and is particularly beneficial for drugs with narrow therapeutic margin and variability in response. Pharmacogenetics (2,3) studies the proportion of interindividual variability in drug

response that can be explained by genetic variation, investigating specifically the link between the genotype and the pharmacokinetic (PK)/pharmacodynamic (PD) (4) phenotype. In PK/PD studies, phenotypes can be observed (e.g. residual concentrations, biomarker measurements) or estimated (e.g. total exposure or specific PK parameters).

Estimated phenotypes in PK are mainly derived through two methods. The noncompartmental analysis (NCA) (5) calculates for each subject measurements of PK exposure such as the area under the curve (AUC) or maximal and trough-observed concentrations through model-free approaches. It requires rich and balanced data per individual, which makes it inappropriate for studies in advanced phases of drug development or in special populations (paediatrics...). Model-based approaches on the other hand describe the dynamic phenomena through a mathematical model and estimate primary PK parameters summarising physiological processes. These methods involve nonlinear mixed effects models (NLMEM) (6), which jointly analyse data obtained on a set of individuals to determine the typical model parameters (fixed effects), and the parameters of inter and intraindividual variability (random effects), as well as the residual error. NLMEM are better suited to sparse and unbalanced data, and clinical studies can be combined to increase the power to detect genetic effect (7).

**Electronic supplementary material** The online version of this article (doi:10.1208/s12248-015-9726-8) contains supplementary material, which is available to authorized users.

<sup>1</sup> INSERM, IAME, UMR 1137, Faculté de médecine Paris Diderot Paris 7 - site Bichat, 16 rue Henri Huchard, 75018, Paris, France.

<sup>2</sup> Université Paris Diderot, IAME, UMR 1137, Sorbonne Paris Cité, 75018, Paris, France.

<sup>3</sup> Division of Clinical Pharmacokinetics and Pharmacometrics, Institut de Recherches Internationales Servier, Suresnes, France.

<sup>4</sup> University College London, Genetics Institute, London, UK.

<sup>5</sup> INSERM CIC 1414, Université Rennes 1, Rennes, France.

<sup>6</sup> To whom correspondence should be addressed. (e-mail: adrien.tessier@inserm.fr)

**ABBREVIATIONS:** FWER, Family wise error rate; LOQ, Limit of quantification; NCA, Noncompartmental analysis; NLMEM, Nonlinear mixed effects model; PK, Pharmacokinetics; SNP, Single nucleotide polymorphism;  $\alpha$ , Type I error;  $\beta$ , Effect size coefficient;  $\lambda$ ,  $\gamma$ ,  $\xi$ , Penalisation parameters.

Unlike monogenic diseases, caused by mutation of a single gene, the variability in PK/PD is usually the product of a set of markers with low to intermediate effect sizes. The screening of a large number of genetic markers, such as single nucleotide polymorphisms (SNP), is possible thanks to the development of genotyping methods and microarrays (8). This has two major consequences: the number of genetic variants studied tends to be larger than the number of subjects and some of these variants are correlated due to linkage disequilibrium (9).

An exhaustive literature review about clinical pharmacogenetic studies was performed to gain an overview of the methods currently used in this type of study (see the [supplementary materials](#) for the literature review methods we used). Information about the phenotypes (obtained by NCA or modelling), the design of studies, the genetic data and the statistical analysis have been extracted from each publication, and summarised through descriptive statistics. During the period 2010–2012, on 85 pharmacogenetic studies using PK parameters as phenotypes, 69% used NCA and 31% used modelling-based phenotypes for association analysis. About two thirds of the studies included less than 50 subjects and only 15% more than 100 subjects. A limited number of genetic covariates were studied (three genes (minimum=1; maximum=45) and eight SNPs (minimum=1; maximum=198) in average). Genes were finely targeted because they are known to be involved in the PK of studied molecules. Finally, association between NCA-based phenotype and polymorphism were mainly investigated using univariate methods (univariate ANOVA, *t* test...) (57%), multivariate methods (multivariate linear regression) (14%) and stepwise or descriptive methods. Model-based phenotypes were explored using mostly stepwise regression (78%), univariate methods applied on individual parameter estimates (7%) or descriptive methods.

So, although health authorities strongly recommend studying the pharmacogenetics of new chemical entities in development (10,11), there is no consensus on analysis methods to explore a large number of polymorphisms in association with PK phenotypes. Specific approaches and statistical tools are required, which must take into account the small amount of PK information provided by each individual in relation to the number of genetic covariates, and be able to detect a signal in a large number of possible relationships. Penalised regression methods analyse all markers simultaneously and use a penalisation function which shrinks most effect coefficients to select a parsimonious set of markers of the phenotype variability. These methods have been especially used in genome-wide association studies (GWAS) to binary outcomes (12) or quantitative traits (13). Here, we evaluate three such methods: ridge regression (14) adapted to include a test of significance for its derived estimates (15), the Lasso (16) and the HyperLasso (17), a generalisation of the Lasso.

In the present study, we propose to compare those three penalised regression methods with a stepwise approach through a simulation study, to assess their ability to detect the influence of genetic variables on the PK. We apply them to four possible PK phenotypes (two observed concentrations, AUC estimated by NCA, model parameters estimated by NLMEM). This work is based on collaboration with the pharmaceutical industry (Institut de Recherches Internationales Servier (IRIS)) and thus

focuses specifically on issues of clinical trials for drug development. This also provided a real case study to design the simulations which had interesting features, including a nonlinear absorption, resulting from the drug's physicochemical properties, and a large number of SNPs collected thanks to a specific microarray. This could enable a meaningful comparison of phenotypes in a challenging context.

## MATERIALS AND METHODS

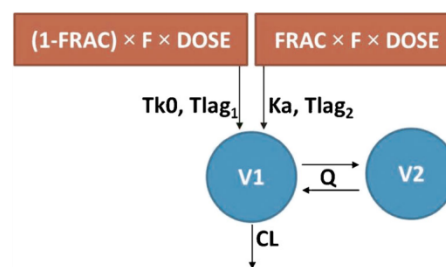
### Overview

In this work, we use a model previously developed to fit the data from a real example for a molecule in early clinical development at IRIS ([Motivating data example](#) section). For this work, we have simulated PK profiles using the model and parameters based on the real data to derive phenotypes ([PK phenotypes](#) section), on which we have applied different association methods ([Statistical methods for genetic association](#) section). We have simulated scenarios ([Simulation study](#) section) for different experimental protocols varying especially the number of subjects and evaluate the methods in terms of detection probability ([Evaluation](#) section).

### Motivating Data Example

The simulation study was inspired by a real dataset which we cannot use for confidentiality issues, but which presented characteristic features encountered during the early phases of drug development, including frequent PK sampling, extensive genetic data, a complex PK profile and a wide range of doses. We start by describing the clinical study and the experimental protocols, as well as the model that will be used in the simulation study.

Drug S developed by IRIS was investigated in three phase I clinical studies performed in a total of 78 adult healthy volunteers. All subjects were genotyped at baseline using a DNA microarray developed by the laboratory. This chip has been developed specifically for PK studies and consists only of SNPs known for being involved in the PK of drugs, *i.e.* markers of phase I and II metabolic enzymes, of SLC or ABC family transports, as well as nuclear receptor genes. The studies included eight dose groups (5, 10, 20, 50, 100, 200, 400 or 800 units) with different allocations (respectively, 6, 6, 24, 12, 12, 6, 6 and 6 subjects by dose) and extensive PK sampling.



**Fig. 1.** Structural PK model of drug S. Double absorption compartments in red and disposition compartments in blue

A two-compartment model with a double absorption function (Fig. 1, Table I) was used to describe the PK of drug S. The nonlinearity with dose was due to the low solubility of drug S, and was modelled through a dependency of the absorption parameters F and FRAC on dose. The relationship of F with dose was expressed through an I<sub>max</sub> model (5) parameterised in I<sub>max<sub>F</sub></sub> and D50<sub>F</sub> while the relationship of FRAC with dose was expressed with an E<sub>max</sub> model (5) parameterised in E<sub>max<sub>FRAC</sub></sub> and D50<sub>FRAC</sub>. T<sub>k0</sub> is the zero order absorption constant rate and T<sub>lag1</sub> the corresponding lag time, K<sub>a</sub> the first order absorption constant rate and T<sub>lag2</sub> the corresponding lag time, V<sub>1</sub> the central compartment volume, V<sub>2</sub> the peripheral compartment volume, Q the intercompartmental clearance and CL the elimination clearance. Interindividual variability was described by an exponential model on F, T<sub>k0</sub>, T<sub>lag2</sub>, V<sub>2</sub>, Q and CL.

### Pharmacokinetic Phenotypes

To capture the PK signal, we considered three types of measures in this study.

#### Observed Concentrations

Observed concentrations are easy to obtain and require no additional analysis step. Here, we considered the last concentration at 192 h after a single dose administration (C<sub>192h</sub>), and the concentration at 24 h (C<sub>24h</sub>), which corresponds to a trough concentration when repeated doses are given during routine treatment.

#### Noncompartmental Approach

AUC was calculated using the linear trapezoidal rule, and extrapolation to infinity was achieved assuming an exponential decay (5,18).

#### Normalisation of Observed and NCA Phenotypes

To take into account the nonlinearity in dose absorption, observed concentrations C<sub>24h</sub> and C<sub>192h</sub>, and AUC obtained by NCA were normalised by dose. An E<sub>max</sub> model (5) of phenotype on dose was fitted for each dataset:

$$\text{Phenotype}_{\text{predicted}_d} = \frac{E_{\text{max}} \times \text{dose}}{ED_{50} + \text{dose}}$$

where  $E_{\text{max}}$  is the maximal value for predicted phenotype,  $ED_{50}$  the dose to obtain 50% of this maximum and dose the administrated amount. Phenotypes predicted by the model at dose  $d$  ( $\text{Phenotype}_{\text{predicted}_d}$ ) and reference dose of 20 units ( $\text{Phenotype}_{\text{predicted}_{20}}$ ) were used to normalise the phenotype observed at dose  $d$  ( $\text{Phenotype}_{\text{observed}_d}$ ) as follow:

$$\text{Phenotype}_{\text{normalised}_d} = \frac{\text{Phenotype}_{\text{predicted}_{20}}}{\text{Phenotype}_{\text{predicted}_d}} \times \text{Phenotype}_{\text{observed}_d}$$

#### Model-Based Approach

Considering N subjects, sampled at one or  $n_i$  times, the concentration  $y_{ij}$  measured in a subject  $i$  receiving a dose  $D_i$  at time  $t_{ij}$ , is described by a nonlinear function (19)  $f$  such as:

$$y_{ij} = f(\theta_i; D_i, t_{ij}) + \varepsilon_{ij}$$

where  $\theta_i$  is the vector of individual parameters assumed to follow a log-normal distribution:

$$\theta_i = \mu e^{\eta_i}$$

with  $\mu$  the population average parameter and  $\eta_i$  the difference between the population average and the individual  $i$  parameter. We assumed that PK parameters follow a log-normal distribution to ensure that they are strictly positives (20).

$\varepsilon_{ij}$  is the residual effect that quantifies the deviation between the model prediction and the measured value for subject  $i$  at time  $j$ . The residual variability was described using a proportional model:

$$g(\theta_i; t_{ij}) = \sigma_{\text{slope}} (f(\theta_i; t_{ij}))$$

where  $\sigma_{\text{slope}}$  represents the proportional component. We typically assume  $\eta_i \sim N(0, \omega^2)$  and  $\varepsilon_{ij} \sim N(0, \sigma^2)$ .

#### Statistical Methods for Genetic Association

We assume that a linear model links the phenotype to the genetic variants as in:

$$\text{Phenotype}_i = \beta_0 + \sum \beta_k \cdot \text{SNP}_{ik} + \varepsilon_i; \text{SNP}_{ik} = \{0, 1, 2\}$$

where  $\text{Phenotype}_i$  is a vector for individual phenotypes,  $\beta_0$  an intercept,  $\text{SNP}_{ik}$  a vector for the genetic variants and  $\beta_k$  a vector for the individual genetic effect size associated to the genetic variant and  $\varepsilon_i$  a residual error following a Gaussian distribution. In this model, the genetic variant takes values 0, 1 or 2, reflecting the number of mutated alleles. All phenotypes are log-transformed to ensure that they follow a normal distribution. To account for type I error inflation due to the multiplicity of tests, all methods used a Sidak correction on the family wise error rate (FWER) to compute a type I error per SNP  $\alpha$ :

$$\alpha = 1 - (1 - \text{FWER})^{\frac{1}{N_t \times P_t}}$$

where  $N_t$  and  $P_t$  are the numbers of SNPs and PK parameters considered simultaneously.

#### Ridge Regression

Ridge regression imposes a penalty on the size of the  $\beta_k$  to reduce the prediction error (15) without preventing the



**Table I.** Population Values ( $\mu$ ) and Interindividual Variability ( $\omega$ ) for Drug S Model (Units Not Disclosed for Confidentiality Reasons)

Parameters		$\mu$	$\omega$ (%)
F <sup>a</sup>	Imax <sub>F</sub>	0.8	32.9
	D50 <sub>F</sub>	41.7	
FRAC <sup>b</sup>	E <sub>max</sub> <sub>FRAC</sub>	0.45	–
	D50 <sub>FRAC</sub>	18.6	
Tlag <sub>1</sub>		0.401	35.1
Tk0		1.59	31.6
Tlag <sub>2</sub>		22.7	–
Ka		0.203	–
V1		1520	–
Q		147	89.9
V2		2130	44.2
CL		94.9	25.1
$\sigma_{\text{slope}}$ (%)		20	–

<sup>a</sup> For doses strictly inferior to 20 units F=1, for doses superior or equal to 20 units F =  $1 - \frac{\text{Imax}_F(\text{dose}-20)}{\text{D50}_F + \text{dose} - 20}$ , where dose is the administrated amount

<sup>b</sup> FRAC =  $\frac{\text{E}_{\text{max}}\text{FRAC} \cdot \text{dose}}{\text{D50}_{\text{FRAC}} + \text{dose}}$ , where dose is the administrated amount

integration of the model variables. We used the approach proposed by Cule *et al.* to set semi-automatically the penalty so that the trace of the projection matrix, which relates the predictions to the observations, is equal to the number of components in a principal component analysis (PCA) of the data. From a Bayesian perspective, this correspond to applying a Gaussian prior of identical variance on the eigenvalues issued from the PCA (21). The ridge regression do not shrunk coefficients estimates to 0. Therefore, we used a Wald test on these coefficients and their standard error (SE), as proposed by Cule *et al.*, (15) to perform the variable selection with the test statistic  $T_0$ :

$$T_0 = \frac{\beta_k}{\text{SE}(\beta_k)}$$

where  $\text{SE}(\beta_k)$  is the standard error of the regression coefficient  $\beta_k$ . Under the null hypothesis,  $T_0$  follows a Student  $t$  distribution with a significant threshold equal to the type I error per SNP  $\alpha$ .

#### Lasso

Lasso (16) also uses a penalty function, which from a Bayesian point of view corresponds to using a double-exponential (DE) probability density as a prior on  $\beta_k$ . The Lasso sets some coefficients to 0 for sufficiently large values of the tuning parameter; this allows variable selection and ensures a parsimonious model. The regularisation parameter  $\xi$  is calculated to achieve a target FWER, using the following expression (17):

$$\xi = \phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \sqrt{\frac{N}{\sigma^2}}$$

where  $\alpha$  is the type I error per SNP,  $\phi^{-1}$  is the inverse normal distribution function,  $N$  the number of subjects and  $\sigma$  the standard error of the phenotype considered.

#### HyperLasso

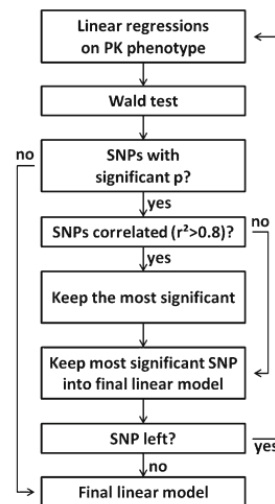
HyperLasso (17) is derived from the Lasso. Here, the penalisation corresponds to using a normal-exponential gamma (NEG) distribution as a prior on  $\beta_k$  and depends on two parameters: a shape parameter  $\lambda$  and a scale parameter  $\gamma$ . The sharp peak at zero and the flatter tail of the NEG distribution favour sparse solutions but the estimates of larger effects are shrunk less severely than the Lasso. The smaller the shape parameter, the heavier the tails of the distribution and the more peaked at zero, which can result in fewer correlated SNPs being selected. The shape parameter  $\lambda$  was set to 1 in our study, which gives realistic effect size distributions (22). As for Lasso, the scale parameter  $\gamma$  is calculated, again depending on a type I error per SNP  $\alpha$ , using the following expression (17):

$$\frac{\text{sign}(\beta)(2\lambda + 1)}{\gamma} \frac{D_{-(2\lambda+2)}\left(\frac{|\beta|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta|}{\gamma}\right)} = \phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \sqrt{\frac{N}{\sigma^2}}$$

where  $\alpha$  is the type I error per SNP and  $D$  the parabolic cylinder function (23).

#### Stepwise Procedure

We use the algorithm in Fig. 2 inspired by Lehr *et al.* (24) and based on univariate regression: (i) PK phenotypes are regressed on each SNP and a Wald test is applied with a significance threshold equal to  $\alpha$ . To account for linkage disequilibrium, among selected SNPs showing strong correlation ( $r^2 > 0.8$ ), only the most significant is kept. Then (ii) the most significant among selected SNPs is included in the linear regression of the PK phenotype on the SNPs. These iterative steps are performed until no more SNP enter the linear model.



**Fig. 2.** Stepwise procedure algorithm, adapted from Lehr *et al.* (24)

### Simulation Study

#### Genotypes

SNPs were simulated using the Hapgen2 software (25) based on the DNA microarray used in clinical studies. To simulate genetic variants for the 200 data sets while retaining the correlations between variants found in the human genome, we used a reference panel of Hapmap genotypes data set (Hapmap 3 release 2) for a Caucasian population (26). Hapgen2 simulates genotype with the same LD patterns as the reference data (25).

Summaries of simulated genotypes, *i.e.* minor allele frequencies, Hardy-Weinberg equilibrium and LD plots, can be found in supplementary file (Simulated polymorphisms information).

#### Pharmacokinetic Profiles

Using the model and parameters estimates from the motivating data example (Table I), we simulated concentrations after a single dose administration according to an extensive sampling schedule with 16 samples per subject at 0.5, 1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 48, 72, 96, 120 and 192 h after taking the tablet. We did not include a limit of quantification (LOQ), and the data were simulated without censoring. NCA method was applied on these simulated profiles and for the model-based approach, population and individual parameters were re-estimated using the Monolix software (27) and SAEM estimation algorithm (28). Individual CL, Q and V2 were used as PK phenotype in a post-model covariate analysis step.

In each simulation scenario, we performed two simulations: in the first simulation (null hypothesis  $H_0$ ) we assumed there was no effect of genetics on the PK parameters; in the second simulation (alternative hypothesis  $H_1$ ), six SNPs were drawn randomly and we assumed they had an impact on the log-transformed clearance according to the following model:

$$\log(CL_i) = \log(\mu_{CL}) + \sum_{k=1}^6 \beta_k \times \text{SNP}_{ik} + \eta_{iCL}$$

where  $CL_i$  is the individual clearance,  $\mu_{CL}$  the population clearance,  $\beta_k$  the effect size associated to the genotype  $\text{SNP}_{ik}$  and  $\eta_{iCL}$  the interindividual variability in clearance.

For each SNP, the associated effect size  $\beta_k$  is computed as a function of the coefficient of genetic component ( $R_{GC_k}$ ) and the minor allele fraction ( $p_k$ ), according to the following equation:

$$\beta_k = \sqrt{\frac{R_{GC_k} \times \omega_{CL}^2}{2p_k(1-p_k) - R_{GC_k} \times 2p_k(1-p_k)}}$$

where  $R_{GC_k}$  is the part of the interindividual variability in CL explained by the SNP (expressed in %) and  $\omega_{CL}^2$  is the variance of random effects on CL due to non-genetic sources.

To simulate realistic genetic effects, the six SNPs altogether explained a total  $R_{GC}$  of 30% with unbalanced effect sizes ( $R_{GC_k}$ , respectively, equal to 1, 2, 3, 5, 7 and 12%). These effect sizes were chosen to be consistent with genetic effects observed in clinical studies. For example, warfarin

doses have been associated with three genetic variants in the cytochrome P450 warfarin-metabolising genes CYP4F2 and CYP2C9 and in the warfarin drug target VKORC1, explaining, respectively, 1.5, 12 and 30% of variability (29).

#### Simulation Scenarios

We simulated different scenarios. In the first scenario,  $S_{\text{real}}$ , the design was chosen to be close to the drug S Phase I clinical trials protocol: 78 subjects ( $N$ ) receiving 8 different doses (5, 10, 20, 50, 100, 200, 400 or 800 units) with a similar allocation ratio and 16 sampled times ( $n$ ) at 0.5, 1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 48, 72, 96, 120 and 192 h. A second scenario  $S_{\text{large}}$  using the same design but with more subjects ( $N=384$ ,  $n=16$ ) was considered as well to investigate a large sample situation. This scenario represents an ideal case from a design perspective, with a large number of subjects and a rich protocol for each subject. Figure 3 represents the PK profiles from a simulated data set with  $S_{\text{large}}$  design as function of the genotypes of the SNP explaining 12% of the clearance interindividual variability (under the alternative hypothesis  $H_1$ ). The profiles show an important interindividual variability. As expected, last concentrations are lower in heterozygotes and rare homozygotes under  $H_1$ , due to the increase of clearances.

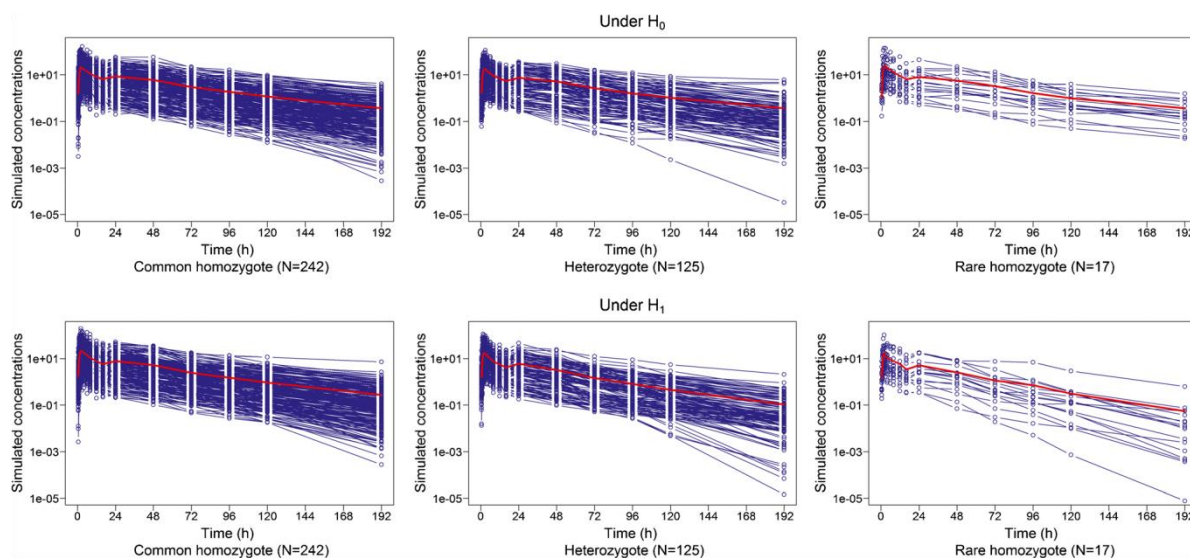
To evaluate the effect of the PK model structure including the nonlinearity of absorption, three additional scenarios were simulated using the same design than  $S_{\text{large}}$  but different structural models: a nonlinear absorption without interindividual variability on F ( $S_{\text{large, no IIV}_F}$ ), a linear PK ( $S_{\text{linearPK}}$ :  $\text{FRAC}=0.5$ ,  $F=0.8$ ,  $\omega(F)=32.9\%$ ) and a linear PK without variability on F ( $S_{\text{linearPK, no IIV}_F}$ :  $\text{FRAC}=0.5$ ,  $F=0.8$ ,  $\omega(F)=0$ ).

#### Evaluation

Each method, ridge regression, Lasso, HyperLasso and the stepwise procedure, was applied to each PK phenotype: observed C24h and C192h, AUC estimated by NCA and model parameters CL, V and Q estimated by NLMEM. To enable proper power comparison, the target FWER was set to 20% (with a prediction interval for 200 data sets equal to [14.5–25.5]). Under  $H_0$ , an empirical FWER was estimated as the percentage of data sets where at least one SNP was found significant. If necessary, the type I error per SNP  $\alpha$  were corrected empirically until the FWER estimate was not significantly different from 20%. Under  $H_1$ , we recorded for each method and phenotype the number of true positives (TP, corresponding to the selection of a SNP which was indeed associated to CL in the simulation, its maximum over the 200 simulations being 1200) and false positives (FP, corresponding to a SNP selected in the model but not present in the simulation). A 95% confidence interval was also estimated assuming the number of TP or FP follows a Poisson distribution. The true positive rate (TPR) and the false positive rate (FPR) were calculated as follow:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$



**Fig. 3.** Individual concentration *versus* time profiles (blue lines) and mean profiles (red lines) for one simulated dataset under  $H_0$  (top) and under  $H_1$  (bottom), in scenario  $S_{large}$ . The profiles are plotted in log-scale for the Y-axis. The three panels show the profiles for common homozygotes (left), heterozygotes (middle), and rare homozygotes (right)

where FN is the count of false negatives and TN the count of true negatives. The TPR was the main outcome on which we statistically compared the different methods and phenotypes. First, we compared the TPR across the different phenotypes for each method, using Cochran's  $Q$  test (30) in the following way: for each dataset and each phenotype, we defined a variable  $X$  with value 1 for the phenotype(s) with the maximal TPR, and 0 for the other ones; this was done separately for each method. A global test for each method was performed first, with a significance threshold set to 5%; if the phenotypes were found to be significantly different, pairwise comparisons were then performed with a Wilcoxon test, using a Sidak correction for the number of tests (a corrected threshold of 0.009 for six pairwise comparisons).

Once the phenotype yielding the highest TPR across the different methods has been selected, the same approach was applied to compare methods on this phenotype. In each simulation, we also recorded the number of datasets for which all methods detected the same number of SNPs ( $X=1$  for all methods), and the number of datasets for which all methods failed to detect at least one SNP ( $X=0$  for all methods).

The probability to detect a given number  $x$  ( $x=1, \dots, 6$ ) of the six SNPs which were associated to CL was computed as the percentage of data sets simulated under  $H_1$  where  $x$  or more SNPs are selected. Of note, for the model-based analysis, associations were explored on parameters CL, Q and V2. TP were causal variants associated to CL and FP were non-causal variants associated to CL and any variants associated to Q and V2. The FPR for the model-based phenotype designed by CL were then computed by taking into account the FP on any of the three model parameters.

## RESULTS

### Scenario 1: $S_{real}$

#### Control of FWER Under $H_0$

Table II shows the estimates of the empirical FWER under  $H_0$ . All methods tended to be too conservative since the FWER was lower than expected for all methods and phenotypes, except on the last concentration parameter C192h. After an empirical correction of thresholds or penalisation parameters depending on the method, FWER was properly controlled around 20%, as shown in Table II. This correction was applied in the corresponding simulation under  $H_1$ .

To investigate why the FWER was not controlled under  $H_0$  despite the calibration step within each method, we simulated a scenario  $S_{independent}$  similar in design to  $S_{real}$  but with independent SNPs. The FWER estimates in  $S_{independent}$  were non-significantly different from 20% (results in supplementary materials). This suggests that correlations between SNPs lead to a decrease in FWER.

#### Performance Under $H_1$

In each scenario, the total number of SNP associated with the PK was 1200, but only a small number was effectively detected, as shown in Table III. This was even more apparent for associations based on NCA parameter or observed concentrations. The TPR was significantly different between the different phenotypes ( $p < 0.001$  for each of the four methods, according to Cochran's  $Q$  test) and was highest for CL in each two by two phenotype tests ( $p < 0.002$  for all pairwise comparisons). The FPR on the other hand was similar for all phenotypes (Fig. 4a). The methods were then



**Table II.** Empirical Estimates of Family Wise Error Rate Under  $H_0$  for  $S_{\text{real}}$  Scenario

Method		FWER (%)			
		C24h	C192h	AUC	CL
Ridge regression	Without correction <sup>a</sup>	13	21	15.5	9.5
Lasso	Without correction <sup>a</sup>	12	20.5	11.5	12
HyperLasso	Without correction <sup>a</sup>	10.5	17.5	11	11
Stepwise procedure	Without correction <sup>a</sup>	14.5	23.5	14.5	14.5
Ridge regression	After empirical correction <sup>b</sup>	19	21	20.5	19.5
Lasso	After empirical correction <sup>b</sup>	18	20.5	18.5	19.5
HyperLasso	After empirical correction <sup>b</sup>	18	17.5	18	19
Stepwise procedure	After empirical correction <sup>b</sup>	18.5	23.5	19	18

The 95% prediction interval around 20 for 200 simulated data sets is [14.5–25.5]

<sup>a</sup> Set of empirical family wise error rates (FWER) obtained without correction

<sup>b</sup> Set of empirical FWER obtained after correction of thresholds or penalisation parameters

compared for CL, and the TPR was found to be significantly different between the four methods ( $p < 0.001$ ). Ridge regression yielded a higher TPR more often than the other methods ( $p < 0.003$  for all pairwise comparison), while Lasso, HyperLasso and the stepwise procedure were comparable.

The probability to detect at least one genetic variant on CL estimates was low, around 40% for all methods (Fig. 5a). This probability decreased quickly when trying to detect more variants and reached 0 for three variants or more.

#### Scenario 2: $S_{\text{large}}$

##### Control of FWER Under $H_0$

As expected, due to the larger number of subjects, the estimated FWER increased with comparison to the scenario  $S_{\text{real}}$ , but remained below the target of 20% for some methods. Empirically corrected thresholds and penalty terms were determined for each combination of phenotype and method to obtain FWER estimates of 20% and were then used for the simulations under  $H_1$  (results in [supplementary materials](#)).

##### Performance Under $H_1$

The number of TP increased clearly compared to the previous scenario (Table IV). Concerning the comparison between phenotypes, we found similar results as for  $S_{\text{real}}$ : all

methods were more powerful when applied to CL compared to the other phenotypes ( $p < 0.001$  for all pairwise comparison). The number of FP also increased and was quite high, but increased proportionally less than the number of TP for the model-based phenotypes, C192h and AUC. For C24h on the other hand, the number of FP exceeded the number of TP. Thus, for a similar FPR between all phenotypes, the TPR was higher for CL (Fig. 4b). The TPR for ridge regression was higher than Lasso and HyperLasso ( $p < 0.001$  with methods applied to CL). The TPR for the stepwise procedure was intermediate; the 2 by 2 tests were not significant for the comparison of the stepwise procedure *versus* ridge regression or Lasso.

As expected, with increasing the number of subjects, the power of all methods increased to reach almost 100% to detect at least one genetic variant (Fig. 5b). Then, the power decreased when trying to detect more variants. Departure in methods was observed on the power to detect at least three and more variants, with the ridge regression and stepwise procedure showing higher power.

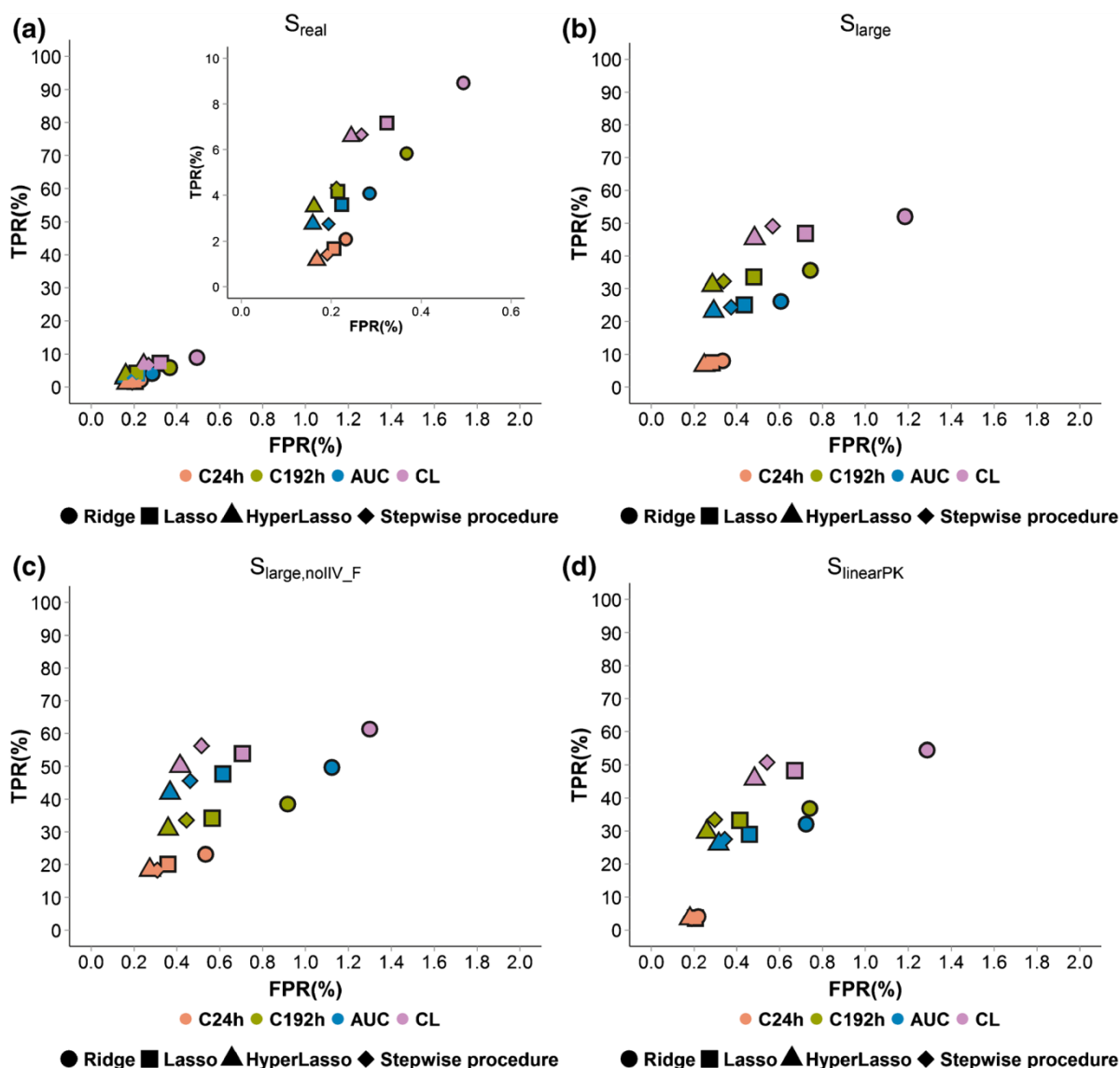
##### Influence of the Structural PK Model Under $H_1$

In both scenarios, regardless of the association method, the power to detect a gene effect was higher using PK parameter obtained by NLMEM than AUC estimated by NCA or observations. We investigated whether this was due to the specific features in the PK model, *i.e.* the nonlinearity

**Table III.** Counts of True and False Positives Under the Alternative Hypothesis for  $S_{\text{real}}$  Scenario

Method		C24h	C192h	AUC	CL	Q	V2
Ridge regression	TP	25 [16;37]	70 [55;88]	49 [36;65]	107 [88;129]	–	–
Lasso	TP	20 [12;31]	50 [37;66]	43 [31;58]	86 [69;106]	–	–
HyperLasso	TP	14 [8;23]	42 [30;57]	33 [23;46]	79 [63;98]	–	–
Stepwise procedure	TP	17 [10;27]	52 [39;68]	33 [23;46]	80 [63;100]	–	–
Ridge regression	FP	79 [63;98]	125 [104;149]	97 [79;118]	97 [79;118]	42 [30;57]	29 [19;42]
Lasso	FP	70 [55;88]	73 [57;92]	76 [60;95]	54 [41;70]	34 [24;48]	22 [14;33]
HyperLasso	FP	57 [43;74]	55 [41;72]	54 [41;70]	32 [22;45]	32 [22;45]	19 [11;30]
Stepwise procedure	FP	65 [50;83]	72 [56;91]	66 [51;84]	36 [25;50]	36 [25;50]	19 [11;30]

Total number of true positives (TP) and false positives (FP) with their 95% confidence interval under the alternative hypothesis. On 200 simulated data sets, overall 1200 SNPs were set to impact clearance (maximum TP number)



**Fig. 4.** Percentage of true positive rate (TPR) versus false positive rate (FPR) (dots). Each PK phenotype is represented by one colour and each association method by one symbol. The  $S_{real}$  scenario is on the top-left with a focus on lower values of TPR and FPR (a), the  $S_{large}$  scenario is on the top-right (b), the  $S_{large, noIV_F}$  scenario is on the bottom-left (c), and the  $S_{linearPK}$  scenario is on the bottom-right (d)

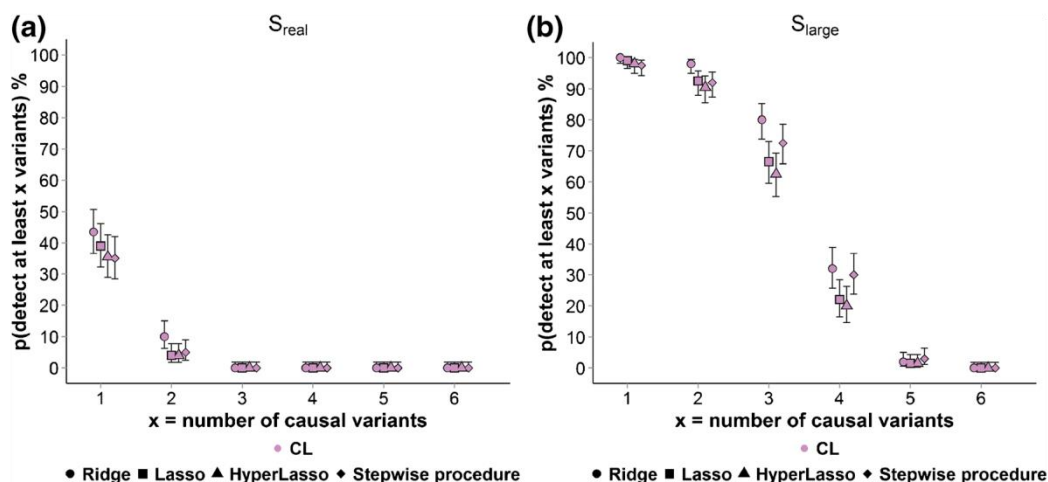
in the absorption model and/or the variability in the bioavailability. When we assumed no variability on the bioavailability parameter  $F$  (scenario  $S_{large, noIV_F}$ ), the difference in the number of TP between CL and AUC was smaller (Fig. 4c), but the TPR for CL remained higher compared to AUC. Assuming a linear absorption, while retaining variability on  $F$ , also reduced this difference (scenario  $S_{linearPK}$ , Fig. 4d) but only when we used a linear absorption model without variability on  $F$  did the benefit of CL over AUC disappear (scenario  $S_{linearPK, noIV_F}$ , results in supplementary materials). Changes in the PK structural model did not affect the number of TP for C192h, while removing the variability on bioavailability increased the number of TP for C24h, although it remained very low.

## DISCUSSION

Many analysis methods have been proposed in the pharmacogenetic literature, depending on the phenotypes studied and the association test. NCA is mainly used to represent the PK exposure, univariate association methods are still widely applied and sample sizes are limited. This work aimed to evaluate 16 combinations of four phenotypes and four methods and the design of the different simulation scenarios was based on actual clinical studies. This realistic setting enabled a meaningful comparison, providing at the same time a challenging context of nonlinear PK and a custom set of polymorphisms.

This work takes place in context of exploratory analyses, and we therefore chose a high FWER for variable selection





**Fig. 5.** Probability estimates (dots) and 95% confidence interval (bars) under  $H_1$  with each method to detect at least  $x$  variants explaining interindividual variability of CL, with  $x$  varying from 1 to 6. The  $S_{\text{real}}$  scenario is on the left (a) and the  $S_{\text{large}}$  scenario is on the right (b)

(20%). The four association methods use estimated phenotypes after an initial estimation step without covariates included in the model. Work on alternative approaches which simultaneously estimate the PK model parameters and the genetic size effects are ongoing (31); they require iterative selection and estimation which increases the computational burden. We decided to study the most common association methods based on a maximum likelihood approach, the ridge regression and Lasso, together with a specific extension for genetic covariates (HyperLasso). Other penalised regression methods have been proposed such as the elastic net method (32), which has shown intermediate performances between ridge regression and Lasso. Other approaches to investigate include modifications of the penalised methods we used, such as the significance test very recently developed for the Lasso (33). In the present work, we did not consider gene-gene and gene-environment interactions. Model-based approaches have been proposed in such contexts, and evaluated on real pharmacogenetic data sets (34). These methods or clustering-based algorithms should be compared to penalised regression methods in simulations close to those presented in this work. Frequentist approaches are most often used in population PK/PD analysis. Bayesian methods (35) may also be worth considering for variable selection but their use remains

limited in population PK/PD studies. The estimate of entire distributions of parameters adds an additional numerical complexity requiring further development beyond the scope of the present work.

Several methodological works associating pharmacogenetics and NLMEM have been published (36–39). Lehr *et al.* have suggested an adaptation of the classical stepwise covariate selection on PK phenotype in NLMEM (24), and a method inspired from Lasso has already been used for the selection of non-genetic covariates in NLMEM (40). But this is the first work comparing model-based approach with NCA in this area. In our study, all methods showed a higher number of TP when used on individual clearances CL from NLMEM, compared to the other phenotypes (AUC, C24h and C192h). Furthermore, relatively to the number of TP, the number of FP was lower for model-based phenotypes (CL, Q and V2) than other phenotypes, also improving the TPR. This finding indicates that using a modelling approach enables better power. Indeed, the modelling approach allows separating the different phases of the PK process (absorption, distribution, metabolism, excretion), improving the interpretation of the genetic effects by the comprehension of mechanisms behind the association between a genetic and a primary PK parameter. The benefits provided by this approach compared to the NCA in

**Table IV.** Counts of True and False Positives Under the Alternative Hypothesis for  $S_{\text{large}}$  Scenario

Method		C24h	C192h	AUC	CL	Q	V2
Ridge regression	TP	96 [78;117]	427 [387;469]	313 [279;350]	624 [576;675]	–	–
Lasso	TP	89 [71;110]	403 [365;444]	301 [268;337]	563 [517;611]	–	–
HyperLasso	TP	80 [63;100]	373 [336;413]	277 [245;312]	545 [500;593]	–	–
Stepwise procedure	TP	91 [73;112]	388 [350;429]	292 [259;327]	590 [543;640]	–	–
Ridge regression	FP	114 [94;137]	253 [223;286]	206 [179;236]	320 [286;357]	32 [22;45]	51 [38;67]
Lasso	FP	97 [79;118]	163 [139;190]	148 [125;174]	159 [135;186]	40 [29;54]	46 [34;61]
HyperLasso	FP	84 [67;104]	97 [79;118]	99 [80;121]	86 [69;106]	36 [25;50]	42 [30;57]
Stepwise procedure	FP	90 [72;111]	115 [95;138]	127 [106;151]	111 [91;134]	39 [28;53]	43 [31;58]

Total number of true positives (TP) and false positives (FP) with their 95% confidence interval under the alternative hypothesis. On 200 simulated data sets, overall 1200 SNPs were set to impact clearance (maximum TP number)

terms of power have also been shown in other areas (41,42), especially when the number of samples per subject is limited, but remained to be demonstrated in the field of pharmacogenetics. In the simulation, we did not include a LOQ and the data were simulated without censoring. In practice however, late measurement of last concentrations could entail a significant number of data below LOQ which would decrease the ability to detect genetic differences. We would expect the NCA approach to be also impacted, as data below the LOQ are usually omitted in NCA, resulting in bias in parameters estimated through such approach (43). In NLMEM on the other hand, different methods based on likelihood have been developed to impute values below the LOQ in NLMEM (44), resulting in unbiased estimates. With such methods, presence of data below the LOQ should not modify the probability to detect genetic effects on phenotypes estimated through NLMEM.

AUC is highly correlated with CL ( $AUC = \frac{Dose}{CL}$ ) when the number of samples per subject is large, as in the design used in simulations. Despite this, the power to detect a gene effect was higher for the model-based approach than for the phenotypes estimated by NCA or observed due to the specific features in the PK model, the nonlinearity in the absorption model and the variability in the bioavailability. Indeed, with a linear absorption and no variability on the bioavailability parameter  $F$  (scenario  $S_{linearPK, noIV_F}$ ), the TPR is similar between CL and AUC (results in supplementary materials). Still in the linear case, an interindividual variability on  $F$  reduces the TPR of AUC (scenario  $S_{linearPK}$ ). In the nonlinear case, although the phenotypes observed or estimated by NCA are normalised, because of estimation errors in the  $E_{max}$  model used for the normalisation, their respective TPR are lower than CL, even with no variability on  $F$  (scenario  $S_{large, noIV_F}$ ). In such a rich design, where the subjects are extensively sampled, AUC is appropriate when the PK is simple (linear, without variability on the absorption process).

In our simulations, we introduced effects from six SNPs on CL parameter in the PK model to simulate concentration profiles under the alternative hypothesis. These settings may favour the model-based phenotypes and also by extension AUC which is directly correlated with CL. With a nonlinear PK model, model-based phenotypes proved much more powerful compared to the phenotype estimated by NCA, even correcting the AUC by dose to take into account the nonlinearity. On the other hand, when the PK model was linear, we found similar results in terms of TPR and FPR for both phenotypes, but only when there was no interindividual variability on bioavailability. Concerning the observed phenotypes, we expected the late concentration C192h to be a good reflection of clearance, and to give similar performance than CL after correction for nonlinearity. However, this was not the case in our results, even with a linear PK. The reason for this is probably that the influence of the other parameters dilutes the impact of the genotypes on this observed phenotype, while AUC only depends on CL. But the results in terms of TP obtain on C192h were better than on C24h. This concentration is not informative for elimination clearance because occurring for many subjects during the rebound simulated using our PK model, so that the poor performance for this phenotype could be expected.

In the model-based approach, using NLMEM individual parameter estimates are derived after the estimation of the population parameters and their precision depends on the amount of individual information available in the data (45).

In the scenarios we simulated, the number of sampling points per subject was large, so that all model parameters, included the phenotype, were well estimated. The number of TP obtained using the simulated clearances (without the estimation step) in both scenarios was only slightly higher than when using the estimated clearance.

The estimates of the probability to detect causal variants were similar between methods in scenario  $S_{real}$ , while a departure was noticed in scenario  $S_{large}$  to detect at least three variants. The ridge regression method exhibited the highest count of TP but to the cost of a higher number of FP. In scenario  $S_{real}$ , the estimates were rather low and fell to nearly 0% for the probability to detect at least three of the six SNPs, indicating a low ability to detect multiple effects in a realistic design. As expected, increasing the number of subjects (scenario  $S_{large}$ ) strongly increased the probability of the different methods to detect more variants. But detecting all the six simulated variants remained very rare, due to the small percentage of variability explained by some variants. Our simulations illustrate the importance of the number of subjects in pharmacogenetic studies: infrequent mutations are unobserved in a small population and association methods are sensitive to the frequency of the variant allele (39).

All methods were fast to run, requiring several minutes to several hours depending on the number of subjects in the scenario. Despite its iterative algorithm, stepwise procedure using univariate regression was the simplest and fastest method in all scenarios. However, the use of a full stepwise procedure using model estimation like proposed by Lehr *et al.* (24) causes a sharp increase in run time (39). Ridge regression and Lasso were intermediate methods in terms of run time and HyperLasso was the slowest method.

There have been few papers comparing association methods in pharmacogenetics applied to PK/PD. The present work complements a previous study by Bertrand and Balding (39), who compared four association methods (ridge regression, Lasso, HyperLasso and a stepwise procedure) on only one kind of PK phenotype, CL estimated by NLMEM. Their setup is close to our scenario  $S_{large}$  in terms of the number of subjects ( $N=300$ ), but each subject was sampled only six times. We compared the number of TP and FP between our studies. Our results for model-based phenotypes are partly different from those presented in this study. Bertrand and Balding found both fewer TP and less FP than in the  $S_{large}$  scenario. These differences result from our respective calculations of TP and FP. In Bertrand and Balding (39), causal variants were removed from the analysis data sets and TP were defined as SNPs correlated with the causal variants with an  $r^2 > 0.05$ . In our study on the other hand, the causal variants were present in the analysis data sets. Using the calculations of Bertrand and Balding, we obtained similar numbers of FP but the numbers of TP remained higher. Bertrand and Balding also explored 1227 SNPs, *i.e.* seven times more SNPs in a similar number of subjects. In the ridge regression, the threshold for the Wald test proposed by Cule *et al.* (15) was therefore much more stringent which could explain that ridge regression detected fewer TP than the other shrinkage-based approaches in their study.

In conclusion, this work shows the critical importance of using modelling approaches for pharmacogenetic studies.



They allow detecting associations between genetic variants and PK more efficiently, in particular in the presence of complex PK involving nonlinearity, where the AUC even corrected for the dose effect was much less sensitive to the genetic effect. In addition, the use of models allows for the analysis of intrinsic parameters with physiological meaning as an elimination or an absorption rate. Our results also reinforce the importance of the number of subjects in pharmacogenetic studies, and suggest that it may not be reasonable to expect to detect even strong genetic effects and/or genetic effect due to rare alleles with small sample sizes. In addition, this work has highlighted statistical difficulties; FWER was not properly controlled and lower than expected. An empirical correction had to be performed to target 20% for the FWER. This decrease was due to the correlation between SNPs, as we showed in an additional simulation with uncorrelated genotypes. Consequently, to enable the comparison across the different methods, the FWER was set empirically for the 16 combinations of methods and phenotypes. A final message from this work is that in our simulation settings, no association method showed a much higher power than the others.

## ACKNOWLEDGMENTS

Adrien Tessier received funding from Institut de Recherches Internationales Servier. The authors thank Laurent Ripoll and Bernard Walther from Institut de Recherches Internationales Servier for their advices in pharmacogenetics. The authors would also like to thank Herve Le Nagard for the use of the computer cluster services hosted on the "Centre de Biomodélisation UMR1137".

## REFERENCES

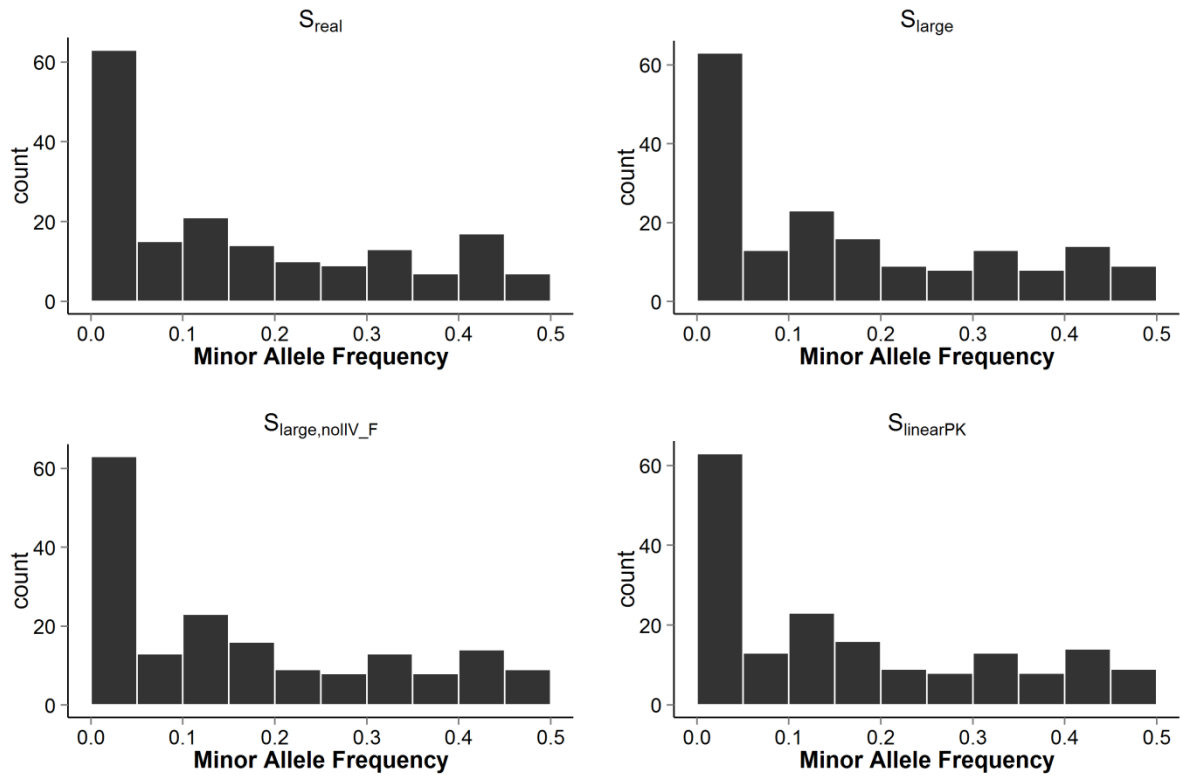
- Aarons L. Population pharmacokinetics: theory and practice. *Br J Clin Pharmacol*. 1991;32(6):669–70.
- Motulsky AG. Drugs and genes. *Ann Intern Med*. 1969;70(6):1269–72.
- Motulsky AG, Qi M. Pharmacogenetics, pharmacogenomics and ecogenetics. *J Zhejiang Univ Sci B*. 2006;7(2):169–70.
- Rowland M, Tozer T. Clinical pharmacokinetics: concepts and applications. Baltimore: Lippincott Williams & Wilkins; 1995.
- Gabrielsson J, Weiner D. Pharmacokinetic and pharmacodynamic data analysis, concepts and applications. 4th ed. Sweden: Swedish Pharmaceutical Press; 2007.
- Sheiner LB, Rosenberg B, Melmon KL. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput Biomed Res Int J*. 1972;5(5):411–59.
- Sheiner LB, Steimer JL. Pharmacokinetic/pharmacodynamic modeling in drug development. *Annu Rev Pharmacol Toxicol*. 2000;40:67–95.
- Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*. 1996;6(7):639–45.
- Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nat Genet*. 2001;29(2):229–32.
- EMA. Guideline on the use of pharmacogenetic methodologies in the pharmacokinetic evaluation of medicinal products. 2012. Report No.: EMA/CHMP/37646/2009.
- FDA. Guidance for Industry and FDA Staff: Pharmacogenetic Tests and Genetic Tests for Heritable Markers. 2007.
- Omoyinmi E, Forabosco P, Hamaoui R, Bryant A, Hinks A, Ursu S, *et al*. Association of the IL-10 gene family locus on chromosome 1 with juvenile idiopathic arthritis (JIA). *PLoS One*. 2012;7(10):e47673.
- Yao T-C, Du G, Han L, Sun Y, Hu D, Yang JJ, *et al*. Genome-wide association study of lung function phenotypes in a founder population. *J Allergy Clin Immunol*. 2014;133(1):248–55.e1-10.
- Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55.
- Cule E, Vineis P, De Iorio M. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*. 2011;12:372.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. 1994;58:267–88.
- Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous analysis of all SNPs in genome-wide and resequencing association studies. *PLoS Genet*. 2008;4(7):e1000130.
- Jaki T, Wolfsegger MJ. Estimation of pharmacokinetic parameters with the R package PK. *Pharm Stat*. 2011;10(3):284–8.
- Dubois A, Bertrand J, Mentre F. Mathematical expressions of the pharmacokinetic and pharmacodynamic models implemented in the PFIM software. 2011. <http://www.pfim.biostat.fr/>.
- Duffull SB, Graham G, Mengersen K, Eccleston J. Evaluation of the pre-posterior distribution of optimized sampling times for the design of pharmacokinetic studies. *J Biopharm Stat*. 2012;22(1):16–29.
- Cule E, De Iorio M. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet Epidemiol*. 2013;37(7):704–14.
- Vignal CM, Bansal AT, Balding DJ. Using penalised logistic regression to fine map HLA variants for rheumatoid arthritis. *Ann Hum Genet*. 2011;75(6):655–64.
- Gradshteyn I, Ryzik I. Tables of integrals, series and products: corrected and enlarged edition. New York: Academic; 1980.
- Lehr T, Schaefer H-G, Staab A. Integration of high-throughput genotyping data into pharmacometric analyses using nonlinear mixed effects modeling. *Pharmacogenet Genomics*. 2010;20(7):442–50.
- Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinforma Oxf Engl*. 2011;27(16):2304–5.
- International HapMap Consortium. The International HapMap Project. *Nature*. 2003;426(6968):789–96.
- Lavielle M, Mesa H, Chatel K. The MONOLIX software. 2010. <http://www.lixoft.eu/>.
- Kuhn E, Lavielle M. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab Stat*. 2004;8:115–31.
- Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, *et al*. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet*. 2009;5(3):e1000433.
- Cochran WG. The comparison of percentages in matched samples. *Biometrika*. 1950;37(3–4):256–66.
- Bertrand J, Bading D, De Iorio M. Penalized regression implementation within the SAEM algorithm to advance high-throughput personalized drug therapy. 22th PAGE Meet Glasg Scotl. 2013;Abstract 2932.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301–20.
- Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. *Ann Stat*. 2014;42(2):413–68.
- Knights J, Chanda P, Sato Y, Kaniwa N, Saito Y, Ueno H, *et al*. Vertical integration of pharmacogenetics in population PK/PD modeling: a novel information theoretic method. *CPT Pharmacomet Syst Pharmacol*. 2013;2(2):e25.
- O'Hara RB, Sillanpaa MJ. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal*. 2009;4(1):85–117.
- Bertrand J, Comets E, Mentre F. Comparison of model-based tests and selection strategies to detect genetic polymorphisms influencing pharmacokinetic parameters. *J Biopharm Stat*. 2008;18(6):1084–102.
- Bertrand J, Comets E, Laffont CM, Chenel M, Mentre F. Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm. *J Pharmacokinet Pharmacodyn*. 2009;36(4):317–39.

38. Bertrand J, Comets E, Chenel M, Mentre F. Some alternatives to asymptotic tests for the analysis of pharmacogenetic data using nonlinear mixed effects models. *Biometrics*. 2012;68(1):146–55.
39. Bertrand J, Balding DJ. Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models. *Pharmacogenet Genomics*. 2013;23(3):167–74.
40. Ribbing J, Nyberg J, Caster O, Jonsson EN. The lasso—a novel method for predictive covariate model building in nonlinear mixed effects models. *J Pharmacokinet Pharmacodyn*. 2007;34(4):485–517.
41. Dubois A, Gsteiger S, Pigeolet E, Mentre F. Bioequivalence tests based on individual estimates using non-compartmental or model-based analyses: evaluation of estimates of sample means and type I error for different designs. *Pharm Res*. 2010;27(1):92–104.
42. Panhard X, Mentre F. Evaluation by simulation of tests based on non-linear mixed-effects models in pharmacokinetic interaction and bioequivalence cross-over trials. *Stat Med*. 2005;24(10):1509–24.
43. Humbert H, Cabiac MD, Barradas J, Gerbeau C. Evaluation of pharmacokinetic studies: is it useful to take into account concentrations below the limit of quantification? *Pharm Res*. 1996;13(6):839–45.
44. Ahn JE, Karlsson MO, Dunne A, Ludden TM. Likelihood based approaches to handling data below the quantification limit using NONMEM VI. *J Pharmacokinet Pharmacodyn*. 2008;35(4):401–21.
45. Savic RM, Karlsson MO. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *AAPS J*. 2009;11(3):558–69.

### 4.3. Matériel supplémentaire

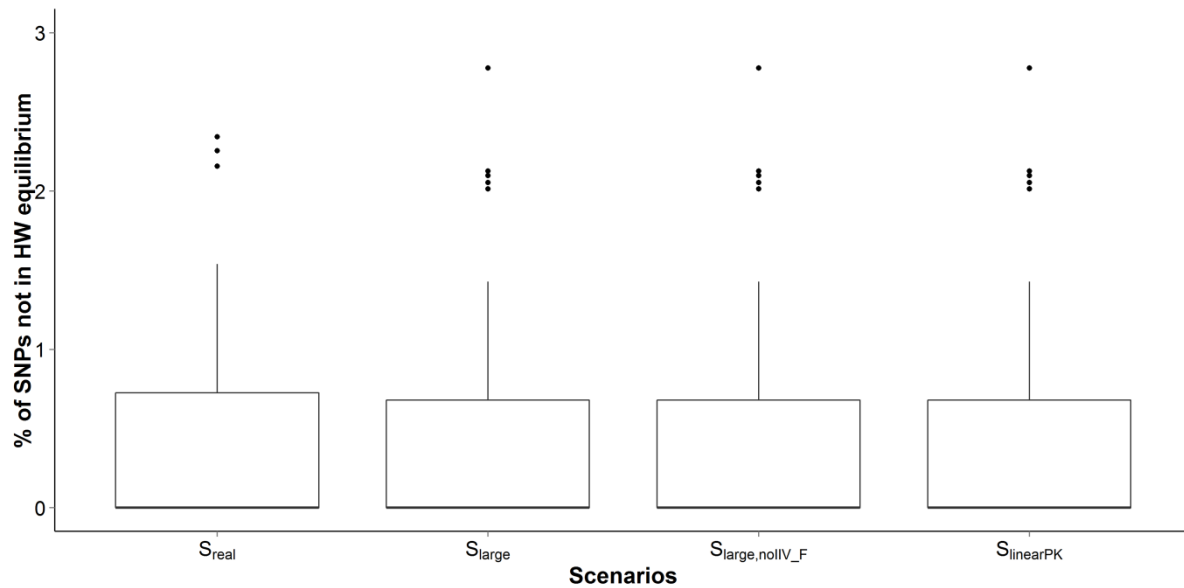
#### SIMULATED POLYMORPHISMS INFORMATION

##### Minor Allele Frequencies



**Fig. Minor Allele Frequency (MAF):** the mean of frequencies of each of the 176 SNPs was computed on the 200 data sets. The distribution of mean MAF is represented for each scenario.

### Hardy-Weinberg Equilibrium



**Fig. Hardy-Weinberg (HW) equilibrium:** the percentage of SNPs not in HW equilibrium was computed in each of the 200 data sets, for each scenario. Departure from HW equilibrium was tested using a  $\chi^2$  test with a Bonferroni adjusted p-value threshold ( $0.05/\text{number of polymorphic SNPs}$ ).

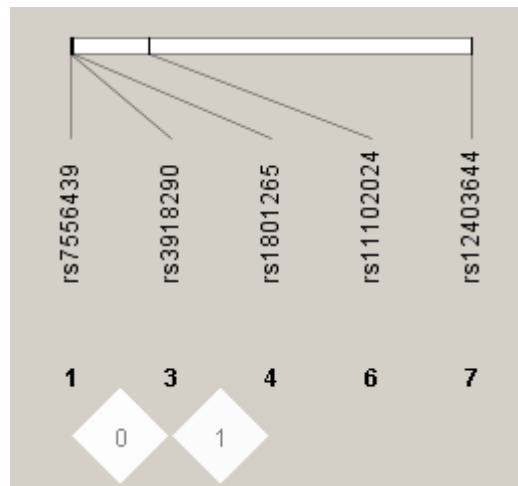
### LD plots

#### Summary

**Table.** Summary on linkage disequilibrium (LD) in an example data set (from scenario  $S_{\text{large}}$ ).

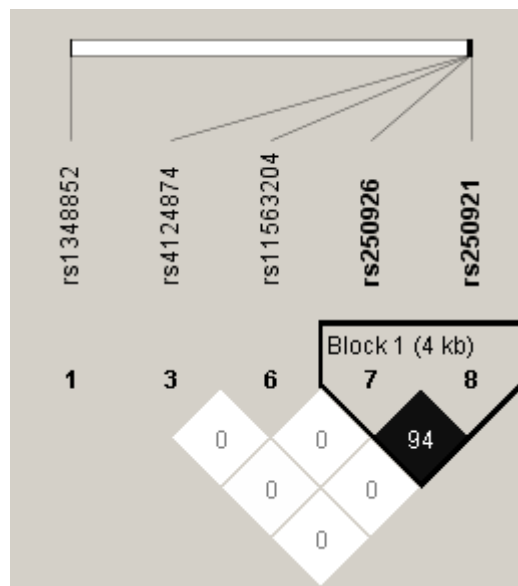
Chromosome	Blocks	Number of SNPs	Number of observed haplotypes
2	1	2	2 (0.807/0.184)
3	2	2	2 (0.667/0.333)
		3	2 (0.534/0.466)
6	3	2	3 (0.711/0.162/0.124)
		2	4 (0.690/0.269/0.030/0.011)
		4	4 (0.742/0.111/0.037/0.108)
7	2	2	3 (0.466/0.456/0.078)
		2	3 (0.454/0.402/0.143)
8	2	5	4 (0.427/0.336/0.167/0.060)
		2	3 (0.363/0.359/0.277)
10	4	4	3 (0.495/0.376/0.128)
		8	6 (0.229/0.310/0.173/0.142/0.090/0.051)
		2	3 (0.094/0.892/0.010)
		2	3 (0.406/0.522/0.072)
11	1	2	3 (0.599/0.303/0.098)
12	1	2	3 (0.465/0.374/0.161)
15	1	2	3 (0.518/0.406/0.076)
19	2	4	4 (0.454/0.306/0.229/0.010)
		3	4 (0.343/0.207/0.338/0.108)
22	1	15	8 (0.238/0.169/0.158/0.107/0.100/0.078/0.56/0.43)

### Chromosome 1



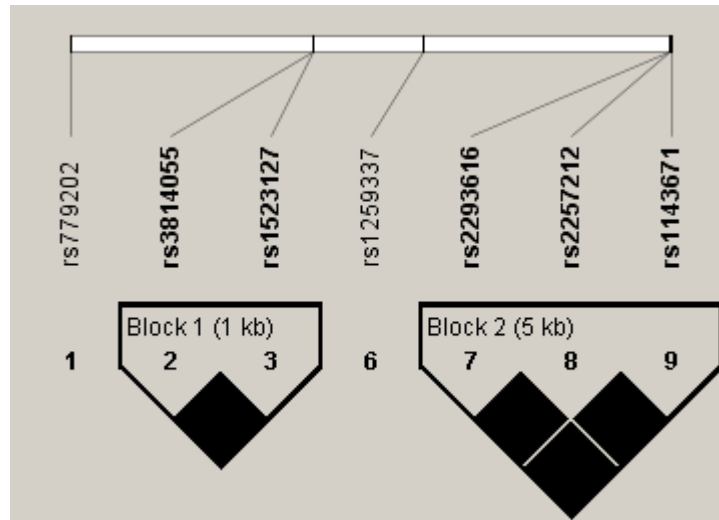
**Fig. LD plot:** Distribution of LD values in chromosome 1 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 2



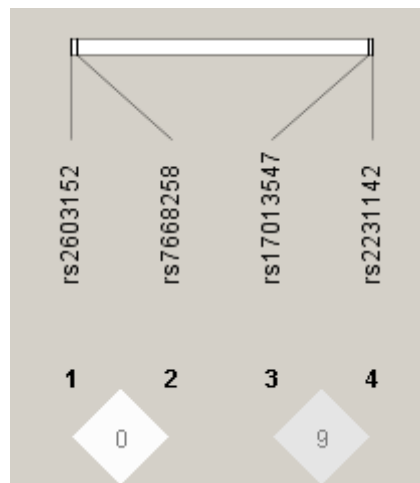
**Fig. LD plot:** Distribution of LD values in chromosome 2 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 3



**Fig. LD plot:** Distribution of LD values in chromosome 3 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD

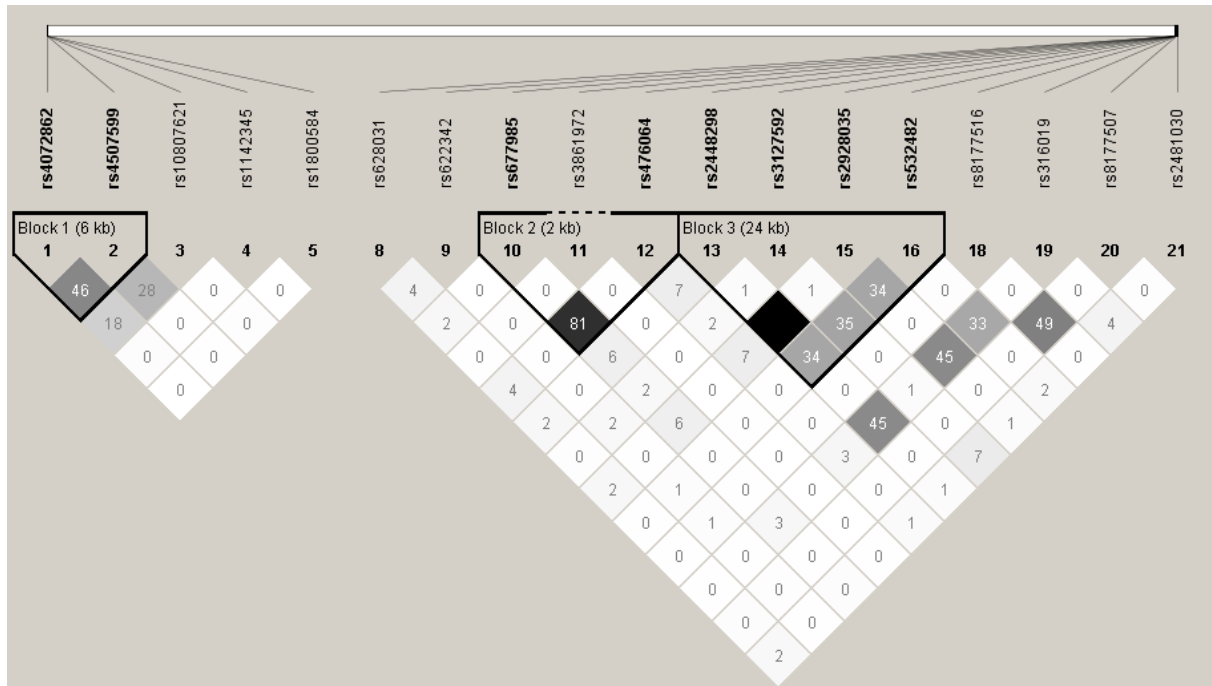
### Chromosome 4



**Fig. LD plot:** Distribution of LD values in chromosome 4 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

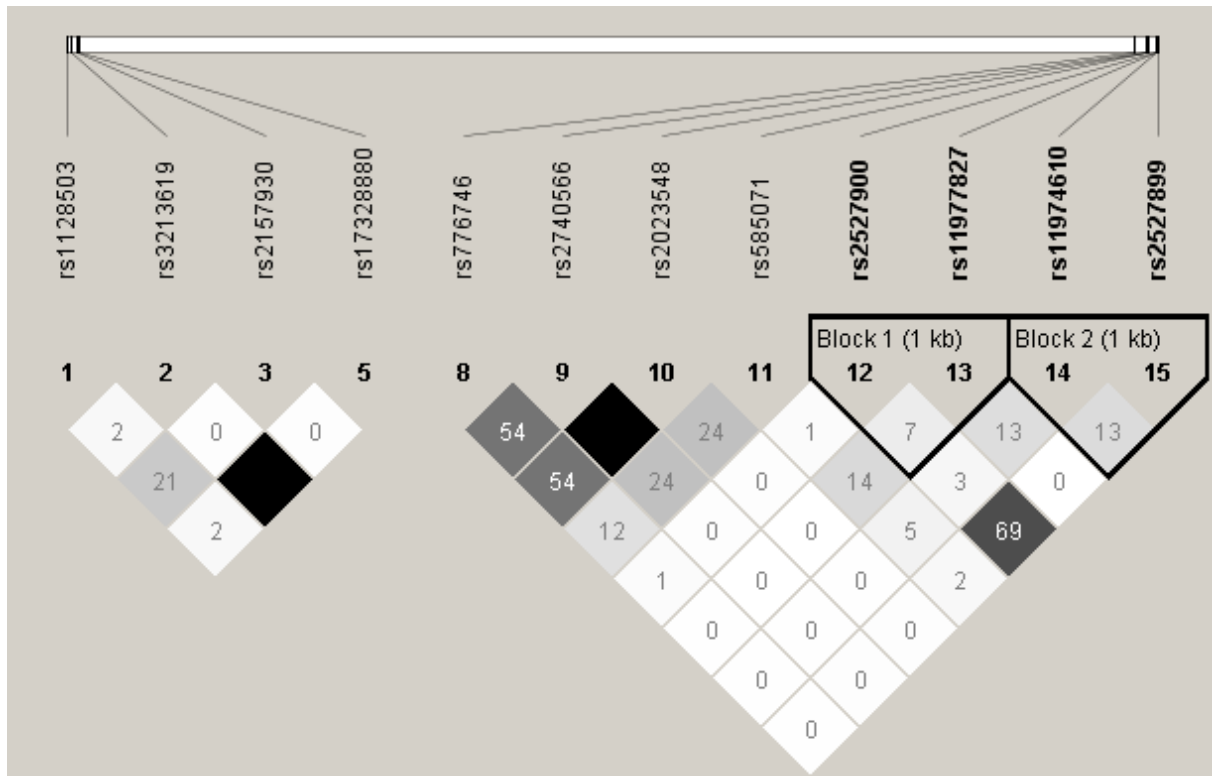


### Chromosome 6



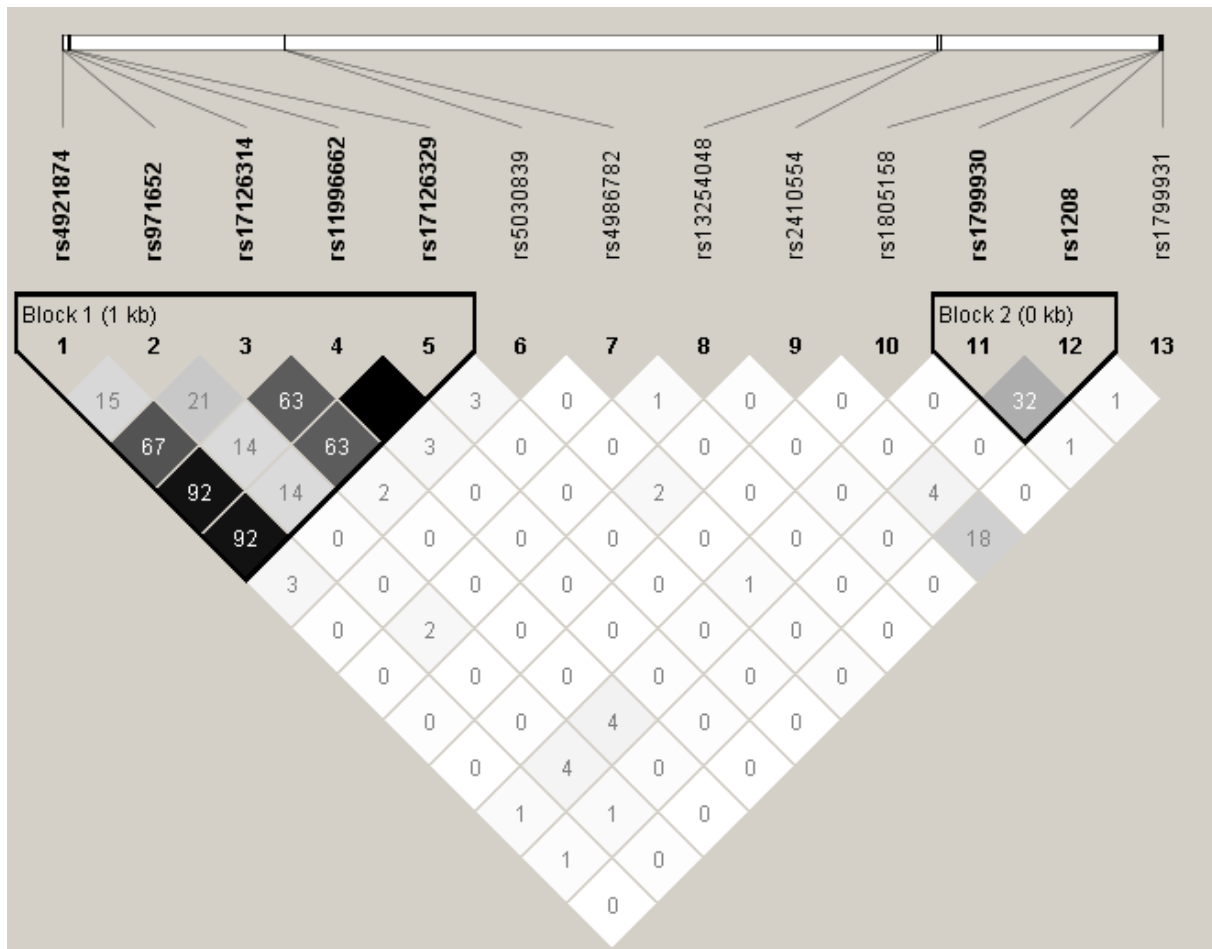
**Fig. LD plot:** Distribution of LD values in chromosome 6 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 7



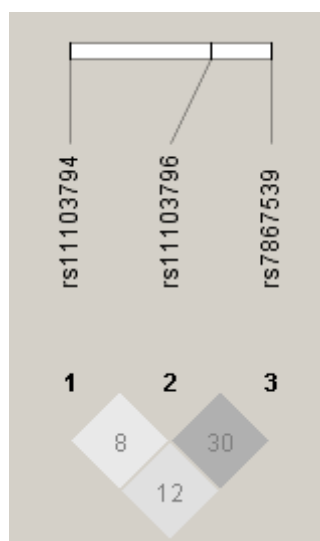
**Fig. LD plot:** Distribution of LD values in chromosome 7 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 8



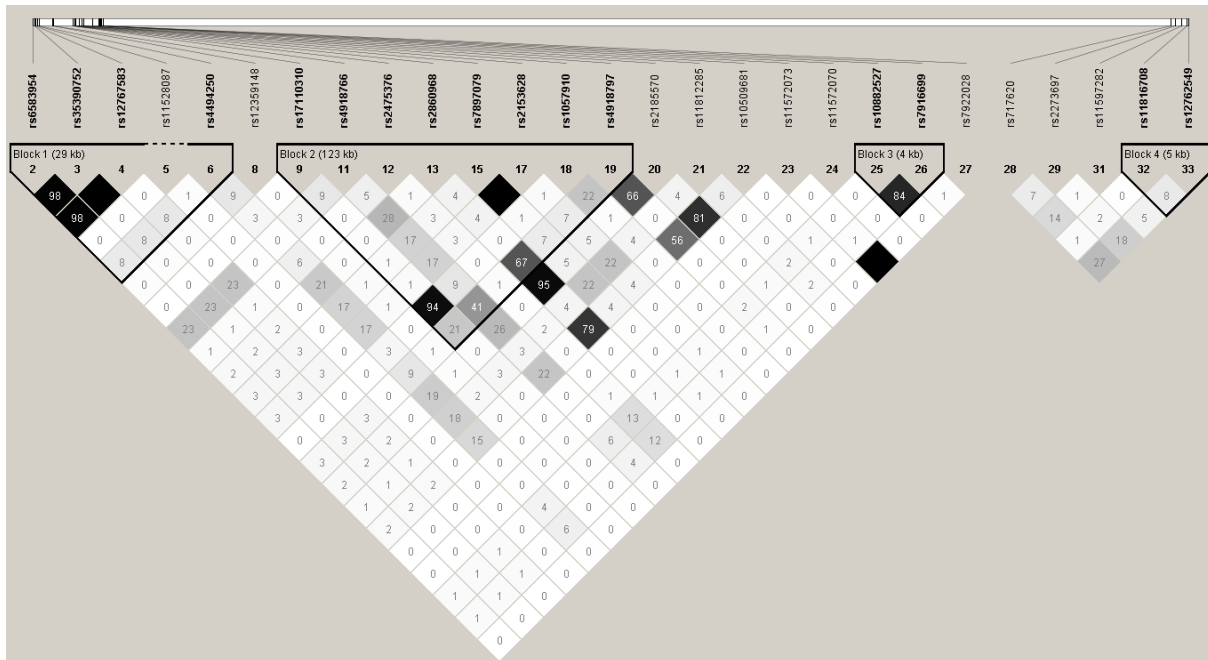
**Fig. LD plot:** Distribution of LD values in chromosome 8 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 9



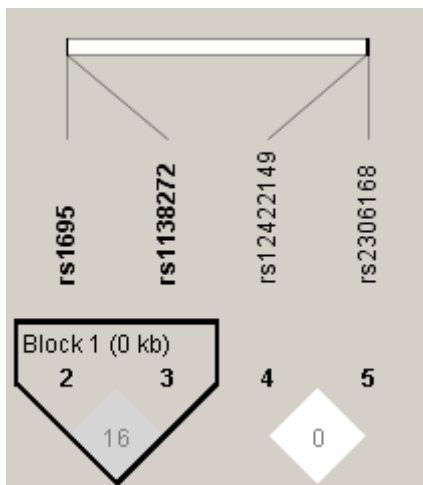
**Fig. LD plot:** Distribution of LD values in chromosome 9 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 10



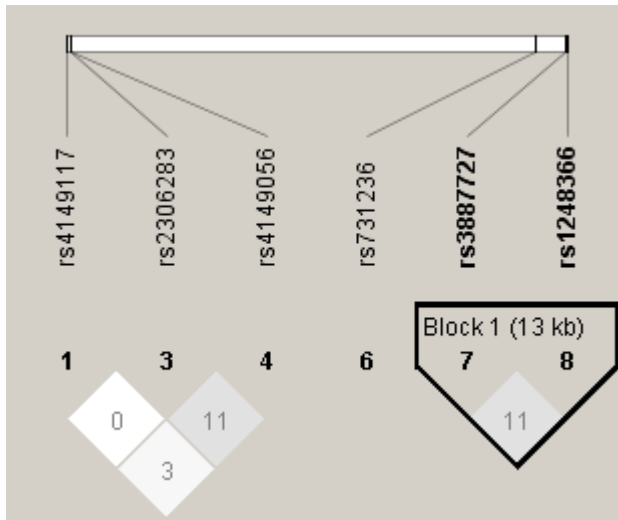
**Fig. LD plot:** Distribution of LD values in chromosome 10 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 11



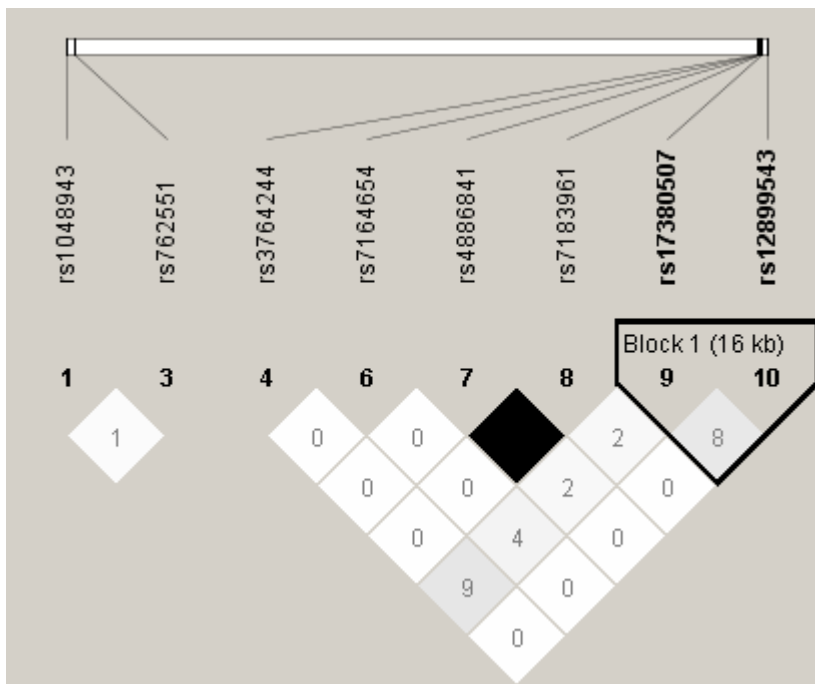
**Fig. LD plot:** Distribution of LD values in chromosome 11 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 12



**Fig. LD plot:** Distribution of LD values in chromosome 12 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 15

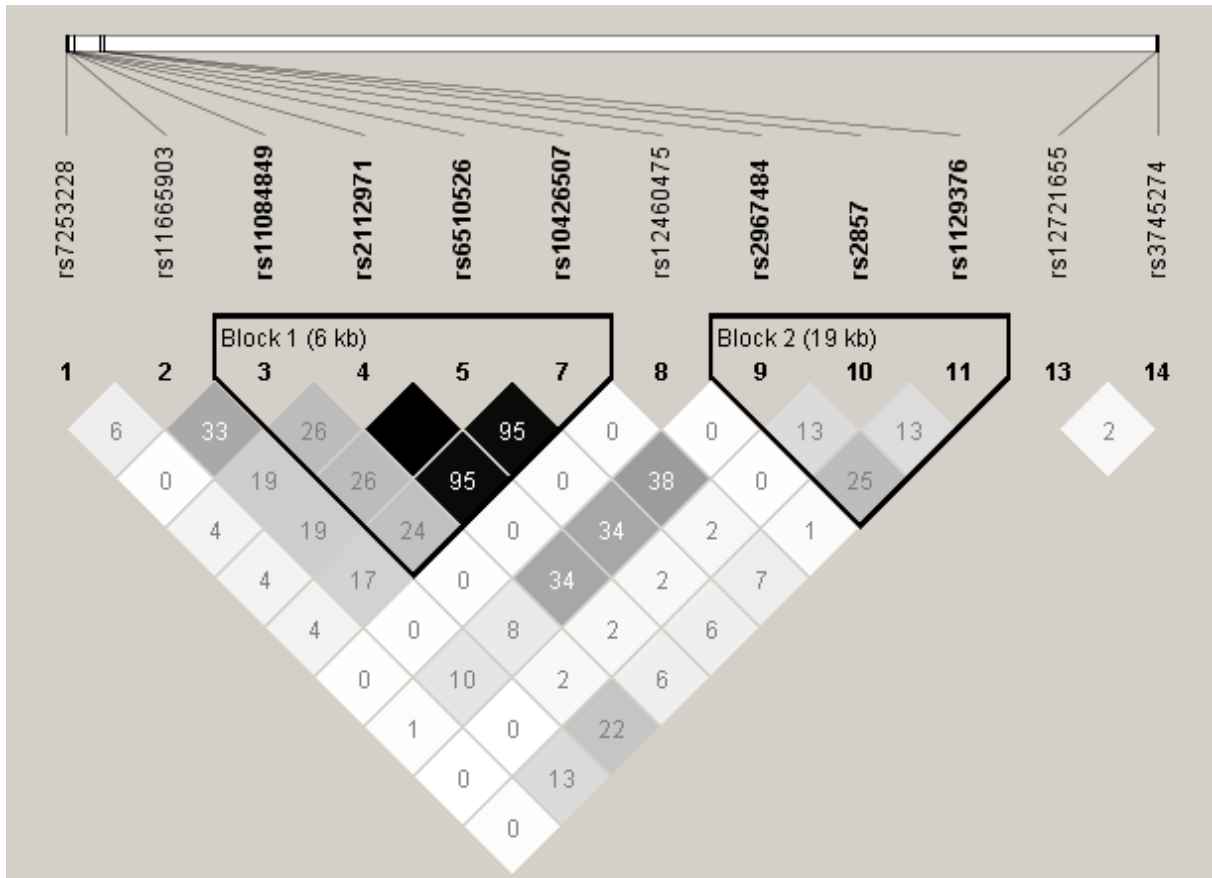


**Fig. LD plot:** Distribution of LD values in chromosome 15 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

### Chromosome 16

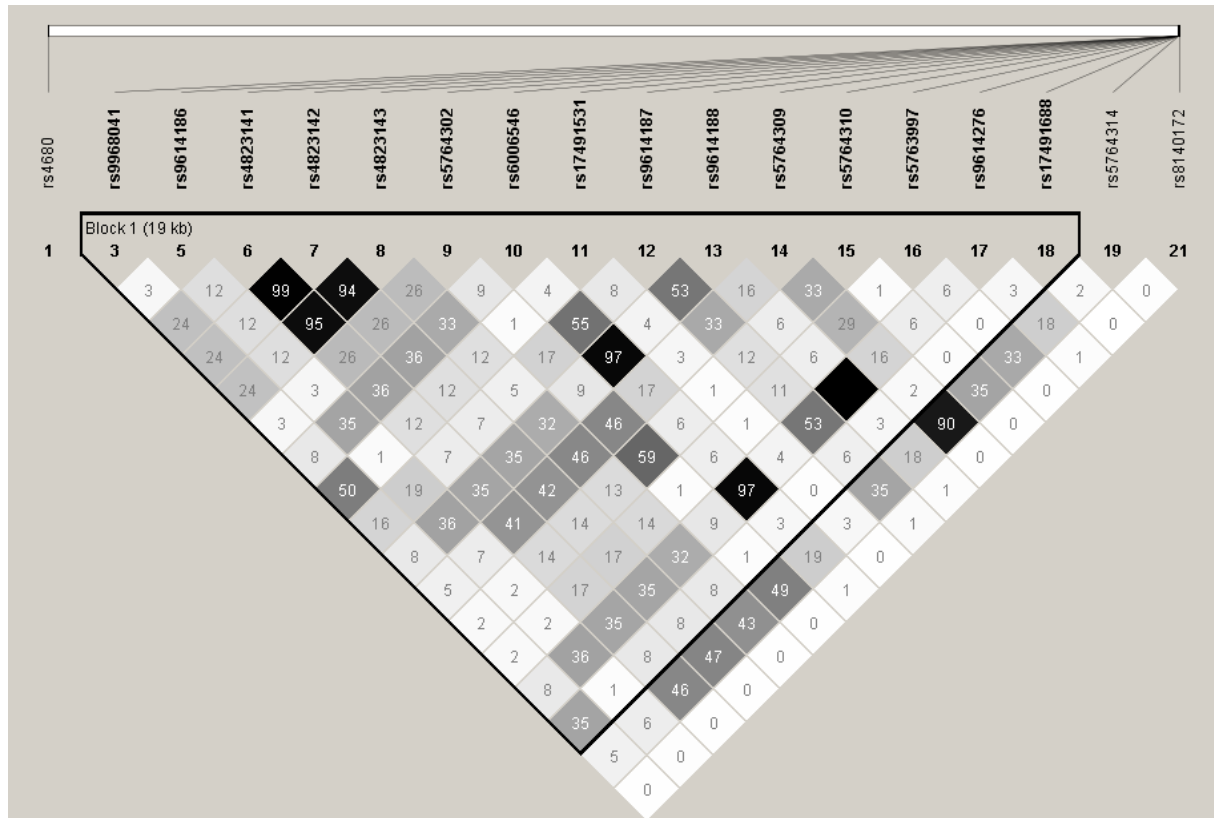
Neither of the two SNPs simulated for this chromosome were polymorphic in the example data set.

Chromosome 19



**Fig. LD plot:** Distribution of LD values in chromosome 19 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

## Chromosome 22

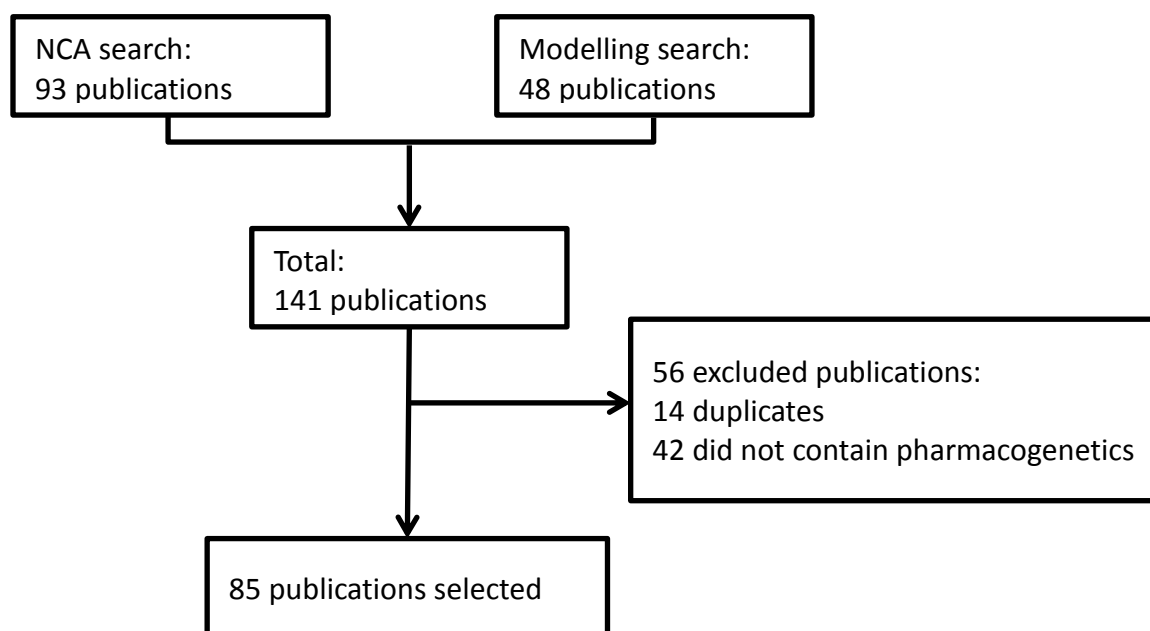


**Fig. LD plot:** Distribution of LD values in chromosome 22 for an example data set (from scenario  $S_{large}$ ). Darker shades indicate higher levels of LD (as measured by  $R^2$ ) and the lighter shades indicate lower levels of LD.

## SUPPLEMENTARY MATERIALS

### Methods for literature review

We performed a bibliographic survey of publications indexed in the Pubmed database to investigate differences in the methods used for pharmacogenetic studies. Selection was limited to publications on clinical trials written in English between January 1<sup>st</sup> 2010 and December 31<sup>th</sup> 2013 for which abstract was available. Two searches based on keywords have been used to retrieve publications dealing with NCA or modelling respectively.



**Fig.** Flow diagram of the literature review.

*NCA search: (((((polymorph\* OR genetic\* OR pharmacogenet\*[Title/Abstract])) AND (pharmacokinet\* AND concentration\*[Title/Abstract])) AND (non-compartment\* OR noncompartment\* OR NCA OR AUC\*[Title/Abstract])) AND Clinical Trial[ptyp] AND hasabstract[text] AND ( "2010/01/01"[PDat] : "2012/12/31"[PDat] ) AND Humans[Mesh] AND English[lang]))*

*Modelling search: (((((polymorph\* OR genetic\* OR pharmacogenet\*[Title/Abstract])) AND (pharmacokinet\* AND concentration\*[Title/Abstract])) AND ("population PK" OR "population pharmacokinetic" OR model\* OR nonlinear OR non-linear OR mixed effect\* OR Nonmem OR Monolix[Title/Abstract])) AND Clinical Trial[ptyp] AND hasabstract[text] AND ( "2010/01/01"[PDat] : "2012/12/31"[PDat] ) AND Humans[Mesh] AND English[lang]))*

### **Softwares**

The C++ software Hapgen version 2.1.2 (1) was used to simulate SNPs from reference haplotype panels provided by the Hapmap Project. Among the 176 SNPs present on the DNA microarray, 55 were also present in the Hapmap reference panel. For the others, we chose the closest variant in the Hapmap database. The SNPs selected from the Hapmap database to target the SNPs from the IRIS DNA microarray were very close with a median of 1730 bases pair departure.

The software Monolix version 4.2.2 (2) is a software using the SAEM algorithm for parameters estimation in nonlinear mixed effects models, which was used to derive the EBE using the PK model and the simulated PK profiles.

The C++ software PLINK version 1.07 (3), is a free and open-source whole genome association analysis toolset, which was used to performed the stepwise procedure.

The statistical software R version 2.15.2 (4) was used to simulate the PK profiles. Some R packages were also used such as “ridge” (5), a package for ridge regression with automatic selection of the regularization parameter.

HyperLasso (6) is a C++ program and was used for Lasso and HyperLasso regression methods. Of note, the phenotype is standardized within this program so  $\sigma$  is equal to 1 in the formula to set  $\gamma$ .



### Results for $S_{\text{large}}$ scenario

Table. Empirical estimates of Family-Wise Error Rate under  $H_0$  for  $S_{\text{large}}$  scenario.

		FWER			
Method		C24h	C192h	AUC	CL
Ridge regression	without correction <sup>a</sup>	13.5	26	19	19
Lasso	without correction <sup>a</sup>	14.5	22	15	16.5
HyperLasso	without correction <sup>a</sup>	14.5	23	16	17.5
Stepwise procedure	without correction <sup>a</sup>	14.5	23	16	17
Ridge regression	after empirical correction <sup>b</sup>	18	18.5	19	19
Lasso	after empirical correction <sup>b</sup>	17.5	22	20	21
HyperLasso	after empirical correction <sup>b</sup>	18	23	20	21
Stepwise procedure	after empirical correction <sup>b</sup>	20.5	23	22.5	21.5

a. Set of empirical family wise error rates (FWER) obtain without correction.

b. Set of empirical FWER obtain after correction of thresholds or penalisation parameters.

The 95% prediction interval around 20 for 200 simulated data sets is [14.5-25.5].

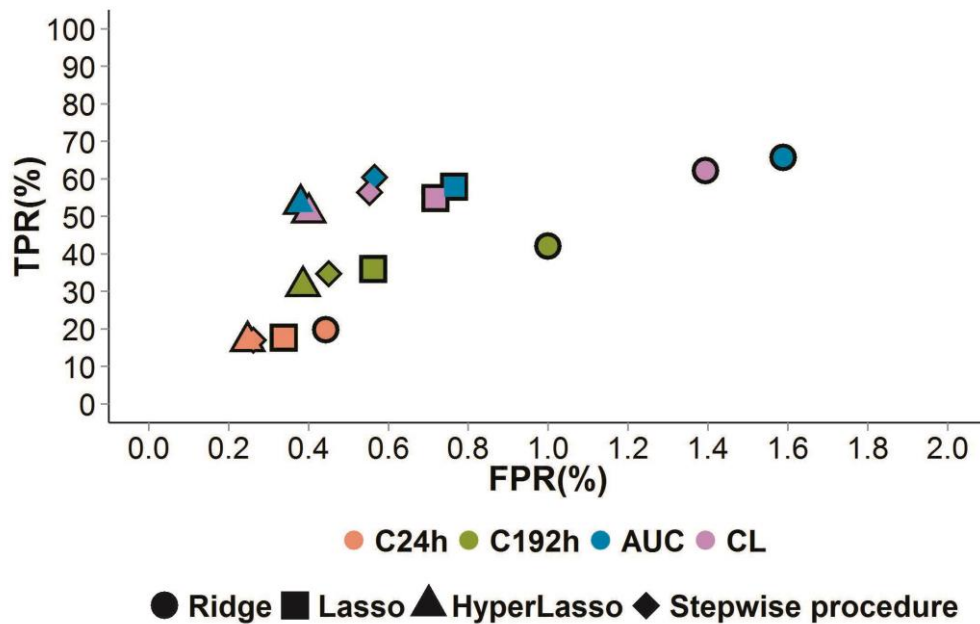
### Results for $S_{\text{independent}}$ scenario

Table. Empirical estimates of Family-Wise Error Rate under  $H_0$  for  $S_{\text{independent}}$  scenario.

		FWER			
Method		C24h	C192h	AUC	CL
Ridge regression		20.5	19.5	16.5	22
Lasso		16.5	17.5	14.5	20
HyperLasso		16.5	17.5	14.5	20
Stepwise procedure		19.5	20	19	21.5

The 95% prediction interval around 20 for 200 simulated data sets is [14.5-25.5].

**Results for  $S_{\text{linearPK, noIV}_F}$  scenario**



**Fig.** Percentage of True Positive Rate (TPR) versus False Positive Rate (FPR) (dots) for scenario  $S_{\text{linearPK}}$  without variability on F. Each PK phenotype is represented by one color and each association method by one symbol.

The remaining difference between TPR for CL and AUC is due to the correction applied on the association methods. The correction depends on the number of tested parameters: one for NCA approach (AUC) and three for model-based approach (CL, Q and  $V_2$ ).

1. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinforma Oxf Engl*. 2011;27(16):2304-2305.
2. Lavielle M, Mesa H, Chatel K. The MONOLIX software. 2010. <http://www.lixoft.eu/>
3. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.
4. R Development Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing. 2013. <http://www.R-project.org/>
5. Cule E, De Iorio M. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet Epidemiol*. 2013;37(7):704-714.
6. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet*. 2008;4(7):e1000130.

# Chapitre 5

## IMPORTANCE DU PROTOCOLE DES ÉTUDES PHARMACOCINÉTIQUES DANS LA PUISSANCE DES ANALYSES PHARMACOGÉNÉTIQUES : INTÉRÊTS DES PROCOTOLES COMBINÉS

### 5.1. Résumé

Le précédent travail de simulation (décrit dans le chapitre 4) montre que l'utilisation de phénotypes PK estimés par MNLEM permet une meilleure puissance de détection des effets génétiques, comparée à l'utilisation d'approches NCA, pour des pharmacocinétiques complexes. Ce travail montre également une performance similaire quant à la détection des variants causaux pour différents tests d'association, une *stepwise procedure* et trois régressions pénalisées (la régression *ridge*, le Lasso et l'HyperLasso). Enfin, les simulations du travail précédant s'inscrivent dans un contexte de phase I du développement clinique et montrent une faible probabilité de détecter des effets génétiques du fait du nombre limité de sujets. En augmentant la taille de l'échantillon, la puissance de détection des tests augmente de façon importante. Mais dans ces simulations tous les sujets ont un protocole de prélèvement riche, ce qui n'est pas réaliste dans le cadre d'essais cliniques. Les protocoles des études de phase II comprennent généralement un nombre important de sujets (100 à 500 patients) mais peu de prélèvements. En effet le nombre de prélèvements peut à l'extrême être réduit à une observation par sujet. Il est démontré que le protocole des études, *i.e.* le nombre de sujets, le nombre de prélèvements par sujet et les temps auxquels les prélèvements sont effectués ont un impact sur la précision d'estimation des paramètres individuels, estimés par une approche bayésienne, à la suite de l'analyse de population par MNLEM. Ainsi, les protocoles de prélèvement épars utilisés en phase II conduisent à des biais dans l'estimation des paramètres individuels qui sont utilisés par la suite dans les analyses de covariables (Savic and Karlsson, 2009). Un protocole peu

informatif peut donc diminuer la puissance de détection des covariables (Combes et al., 2014), notamment génétiques.

Dans ce second travail de simulation, nous cherchons à proposer des protocoles réalisables en phase II, et par extension en phase III, et assurant une puissance raisonnable pour détecter des effets génétiques sur un phénotype PK estimé par approche de population. Au vu des résultats des premiers travaux, seuls deux tests d'association sont évalués dans l'analyse principale (la *stepwise procedure* inspirée de (Lehr et al., 2010) et le Lasso (Tibshirani, 1994)). Les résultats concernant la régression ridge (Cule and De Iorio, 2013) et l'HyperLasso (Hoggart et al., 2008) peuvent être trouvés dans le matériel supplémentaire de ce chapitre.

Dans ce travail de nouvelles études sont simulées à partir du cas réel présenté au chapitre 2, en conservant les mêmes conditions de simulation que précédemment pour la pharmacocinétique et l'effet des polymorphismes génétiques, *i.e.* six variants causaux affectant la clairance. En plus d'une étude de phase I incluant 78 sujets avec un protocole de prélèvement riche (16 observations par sujet), trois études de phase II sont simulées, incluant 306 sujets avec un protocole de prélèvement épars. Une première étude ne comprend qu'une observation par sujet et les deux autres études trois prélèvements dont les temps sont optimisés. L'approche utilisée pour l'optimisation de ces protocoles se base sur la matrice d'information de Fisher, *i.e.* la matrice de variance-covariance des paramètres du modèle, et le programme PFIM ([www.pfim.biostat.fr](http://www.pfim.biostat.fr)) intégrant deux algorithmes d'optimisation permettent d'évaluer rapidement un nombre très important de protocoles (Bazzoli et al., 2010). Au final, quatre scénarios d'analyses sont évalués : un scénario n'incluant que les données issues de l'étude de phase I et les trois autres scénarios combinant les données de phase I et les données issues d'une des études de phase II. L'erreur de type I globale (FWER) est quantifiée et calibrée sous  $H_0$  et la puissance de détection sous  $H_1$  est évaluée. Le  $\eta$ -shrinkage, reflétant la quantité d'information apportée par chaque sujet pour l'estimation des paramètres individuels dans les MNLEM et variant en fonction du protocole de prélèvement, est également estimé (Bertrand et al., 2009; Savic and Karlsson, 2009).

À nouveau les tests d'association s'avèrent plus conservatifs sous  $H_0$ , avant la correction empirique, avec une erreur de type I globale inférieure aux 20% attendus. La puissance de

détection est faible pour le scénario n'incluant que les données de phase I, tandis qu'une augmentation très importante de cette puissance est observée sur les scénarios combinant les données de phase I et de phase II et ce quel que soit le protocole de prélèvement. Ce résultat reflète l'impact attendu du nombre de sujets sur la puissance de détection des effets génétiques. Le fait que cette augmentation de puissance se fasse même avec l'ajout de sujets ayant un protocole de prélèvement très épars est par contre plus surprenant. Cela montre le réel intérêt de prévoir, au sein d'une même étude où seul un protocole de prélèvement épars est possible, un petit groupe de sujets avec un nombre d'observations plus important. Au sein des scénarios combinant les données de phase I et de phase II, des différences de puissance de détection sont tout de même observées, pouvant être reliées au  $\eta$ -shrinkage de chaque protocole de phase II. Le protocole le plus informatif, *i.e.* ayant le  $\eta$ -shrinkage le plus faible, permet la meilleure puissance de détection. La perte du signal génétique est aussi quantifiée pour expliquer ce lien. Ainsi une relation directe entre  $\eta$ -shrinkage, perte du signal génétique et puissance de détection est mise en évidence. Cela reflète le bénéfice de l'utilisation d'approches d'optimisation de protocoles pour la mise en place d'études cliniques pertinentes.

Ce travail fait l'objet d'un article accepté le 11 décembre 2015 dans *CPT: Pharmacometrics and Systems Pharmacology*.

## **5.2. Article 2 (accepté)**

ORIGINAL ARTICLE

AQ2

# Combined Analysis of Phase I and Phase II Data to Enhance the Power of Pharmacogenetic Tests

AQ1

A Tessier<sup>1,2,3\*</sup>, J Bertrand<sup>4</sup>, M Chenel<sup>3</sup> and E Comets<sup>1,2,5</sup>

We show through a simulation study how the joint analysis of data from phase I and phase II studies enhances the power of pharmacogenetic tests in pharmacokinetic (PK) studies. PK profiles were simulated under different designs along with 176 genetic markers. The null scenarios assumed no genetic effect, while under the alternative scenarios, drug clearance was associated with six genetic markers randomly sampled in each simulated dataset. We compared penalized regression Lasso and stepwise procedures to detect the associations between empirical Bayes estimates of clearance, estimated by nonlinear mixed effects models, and genetic variants. Combining data from phase I and phase II studies, even if sparse, increases the power to identify the associations between genetics and PK due to the larger sample size. Design optimization brings a further improvement, and we highlight a direct relationship between  $\eta$ -shrinkage and loss of genetic signal.

CPT Pharmacometrics Syst. Pharmacol. (2015) 00, 00; doi:10.1002/psp4.12054; published online on 0 Month 2015.

## Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC? ☒ Most pharmacogenetic analyses in pharmacokinetic studies recently published included a limited number of subjects (fewer than 50). Previous simulations showed that such sample sizes result in a low probability to detect polymorphisms. But with large numbers of subjects, extensive pharmacokinetic information is difficult to obtain in drug development. • WHAT QUESTION DID THIS STUDY ADDRESS? ☒ This simulation study explored realistic ways to increase the amount of information by combining rich phase I data and sparse phase II data, and optimizing such sparse designs. • WHAT THIS STUDY ADDS TO OUR KNOWLEDGE ☒ This study shows that even sparse data from phase II allow a marked improvement in the probability to detect genetic variants when combined with rich data from phase I, even more when sparse designs are optimized. • HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY AND THERAPEUTICS ☒ The pharmacogenetic analyses should be planned later in drug development to take advantage of larger sample sizes by combining data that would increase the power to detect genetic effects.

Studying the sources of the variability observed in drug response facilitates individualization of prescription. One of the sources of variability in drugs' pharmacokinetics (PK)<sup>1</sup> is the variation in activity of enzymes and transporters involved in the drug absorption, distribution, metabolism, or elimination. Pharmacogenetics<sup>2</sup> studies the genetic component of interindividual variability (IIV) observed in PK to identify populations at risk of treatment inefficacy or adverse effects.<sup>3</sup> Single nucleotide polymorphisms (SNPs) are the genetic variants most frequently studied in pharmacogenetics and screened more and more often in clinical studies.

Genetic data offer some unique challenges, in particular because they may lead to a very unbalanced number of subjects, which impacts the power of tests in pharmacogenetic analyses.<sup>4,5</sup> In a previous simulation work, we showed that typical phase I studies have low power to detect genetic effects because of the limited sample size.<sup>6</sup> On the other hand, phase I studies generally provide good quality PK information, allowing characterization of the PK profile of the drug. We showed that from the different approaches used at this stage to estimate PK parameters, nonlinear

mixed effects models (NLMEM)<sup>7</sup> could be considerably more powerful than noncompartmental analyses (NCA)<sup>8</sup> for complex PK models.<sup>6</sup> Our simulations also showed that increasing the sample size, as in phase II studies, would improve the power to detect genetic variants. However, sparse designs typically used in phase II may result in biased estimations for empirical Bayes estimates (EBE)<sup>9</sup> used in generalized additive model (GAM) covariate analysis procedure.<sup>10</sup>

To increase the detection of genetic covariates, one way could be to combine for the analysis data from a study collected with a rich design, as expected in phase I, with sparser, but still informative, data from a phase II study.

In the present work we propose practical designs involving phase I and phase II data, and we quantify through simulations their ability to detect genetic associations with PK. A motivating example was provided by IRIS (Institut de Recherches Servier), a pharmaceutical industry, to generate realistic genetic and PK data. We compared two association methods, a penalized regression method and a stepwise procedure.<sup>6</sup>

<sup>1</sup>INSERM, IAME, UMR 1137, Paris, France; <sup>2</sup>Université Paris Diderot, IAME, UMR 1137, Sorbonne Paris Cité, Paris, France; <sup>3</sup>Division of Clinical Pharmacokinetics and Pharmacometrics, Institut de Recherches Internationales Servier, Suresnes, France; <sup>4</sup>University College London, Genetics Institute, London, UK; <sup>5</sup>INSERM CIC 1414, Université Rennes 1, Rennes, France. \*Correspondence: A Tessier (adrien.tessier@inserm.fr).

Received 10 August 2015; accepted 11 December 2015; published online on 0 Month 2015. doi:10.1002/psp4.12054



## MATERIALS AND METHODS

### Simulation study

**F1** **Figure 1** presents the framework of the simulation study, which was designed based on PK data from drug S (IRIS) collected in 78 subjects from three phase I clinical studies.<sup>11</sup> All subjects were genotyped at baseline using a DNA microarray developed by IRIS of 176 SNPs known for being involved in the PK of drugs. These 176 polymorphisms were matched to a reference Hapmap panel (Hapmap 3, release 2) for a Caucasian population<sup>12</sup> and we used the Hapgen2 software<sup>13</sup> to simulate genetic variants retaining their frequencies and the correlations between polymorphisms found in the human genome (see details in **Supplementary Material, Supplementary Figures S1–S3**).

**T1** PK profiles were simulated with a two-compartment model with dose-dependent double absorption (**Supplementary Figure S4**), with the parameters in **Table 1**, under two conditions: 1) no gene effect ( $H_0$ ); 2) gene effect on clearance CL ( $H_1$ ). Under  $H_1$ , six SNPs were drawn randomly without physiological assumptions or prior knowledge, and assumed to explain in total 30% of the IIV on CL through the following additive genetic model on the log-transformed CL:

$$\log(CL_{sim_i}) = \log(\mu_{CL}) + \sum_{k=1}^6 \beta_k \times SNP_{ik} + \eta_{iCL} \quad (1)$$

where  $CL_{sim_i}$  is the simulated individual clearance,  $\mu_{CL}$  the typical clearance,  $\beta_k$  the effect size associated to the variant allele of  $SNP_{ik}$ , and  $\eta_{iCL}$  the interindividual random effect for clearance of subject  $i$ . Causal SNPs were different from one dataset to another. Assuming an additive genetic model, genotypes take values 0, 1, or 2, reflecting the number of mutated alleles. We chose this model to simplify the simulations but dominant or recessive genetic models could be easily simulated by changing genotype values.  $\beta_k$  was computed as a function of the coefficient of genetic component ( $R_{GCK}$ , the percentage of the interindividual variability in CL explained by the SNP) and the minor allele fraction ( $p_k$ ), as follows:

$$\beta_k = \sqrt{\frac{R_{GCK} \times \omega_{CL}^2}{2p_k(1-p_k) - R_{GCK} \times 2p_k(1-p_k)}} \quad (2)$$

where  $\omega_{CL}^2$  is the variance of interindividual random effects on CL due to nongenetic sources.  $R_{GCK}$  was respectively equal to 1, 2, 3, 5, 7 and 12% for the 6 causal variants<sup>14</sup> to mimic a multifactorial genetic effect. Then under  $H_0$ ,  $\omega_{CL}^2 = 0.06$  (as in **Table 1**), while under  $H_1$ , 30% of the variance is explained by the genetics so that  $\omega_{CL}^2 = 0.04$  (example on the magnitude of simulated effect sizes is available in **Supplementary Table S1**).

The simulated datasets were then fitted with the base model without genetic covariates. Individual clearance estimates ( $EBE_{CL_i}$ ) were estimated and all associations with the 176 simulated polymorphisms were tested assuming a linear relation without reestimating model parameters, as in a GAM analysis.<sup>10</sup>

We compared two association methods to detect gene effects. Lasso<sup>15</sup> is a multivariate penalized regression which simultaneously estimates effect size coefficients and selects variants by setting a large number of coefficients to 0. The penalty is set by a tuning parameter ( $\lambda$ ) which depends on  $\alpha$ , the type I error per test, and the number of subjects<sup>16,17</sup> (**Figure 1**). Alternatively, in practice the penalty can be determined through permutation or cross-validation methods, which are more time-consuming. A stepwise procedure includes relationships one-by-one depending on the significance of a Wald test compared to a threshold  $\alpha$ . The correlation between two significant SNPs, due to linkage disequilibrium, is computed through the Pearson's correlation coefficient  $r$  and if two significant SNPs are strongly correlated ( $|r| > 0.89$ ), only the most significant is kept. Finally, the most significant variant among selected SNPs is kept in the final model and steps are repeated until no more association is significant<sup>6</sup> (**Figure 1**).

In both approaches, we control the Family Wise Error Rate (FWER), representing the percentage of datasets where at least one variant is selected under  $H_0$ , by correcting the nominal  $\alpha$  by the number of tests performed (Sidak correction) corresponding to the number of polymorphic SNPs considered  $N_t$  (**Figure 1**). The FWER was set to 20% (with a prediction interval for 200 datasets equal to [14.5–25.5]) for an exploratory analysis. The prediction interval determined when to adjust  $\alpha$  to control the FWER under  $H_0$ .

### Simulated designs and analysis scenarios

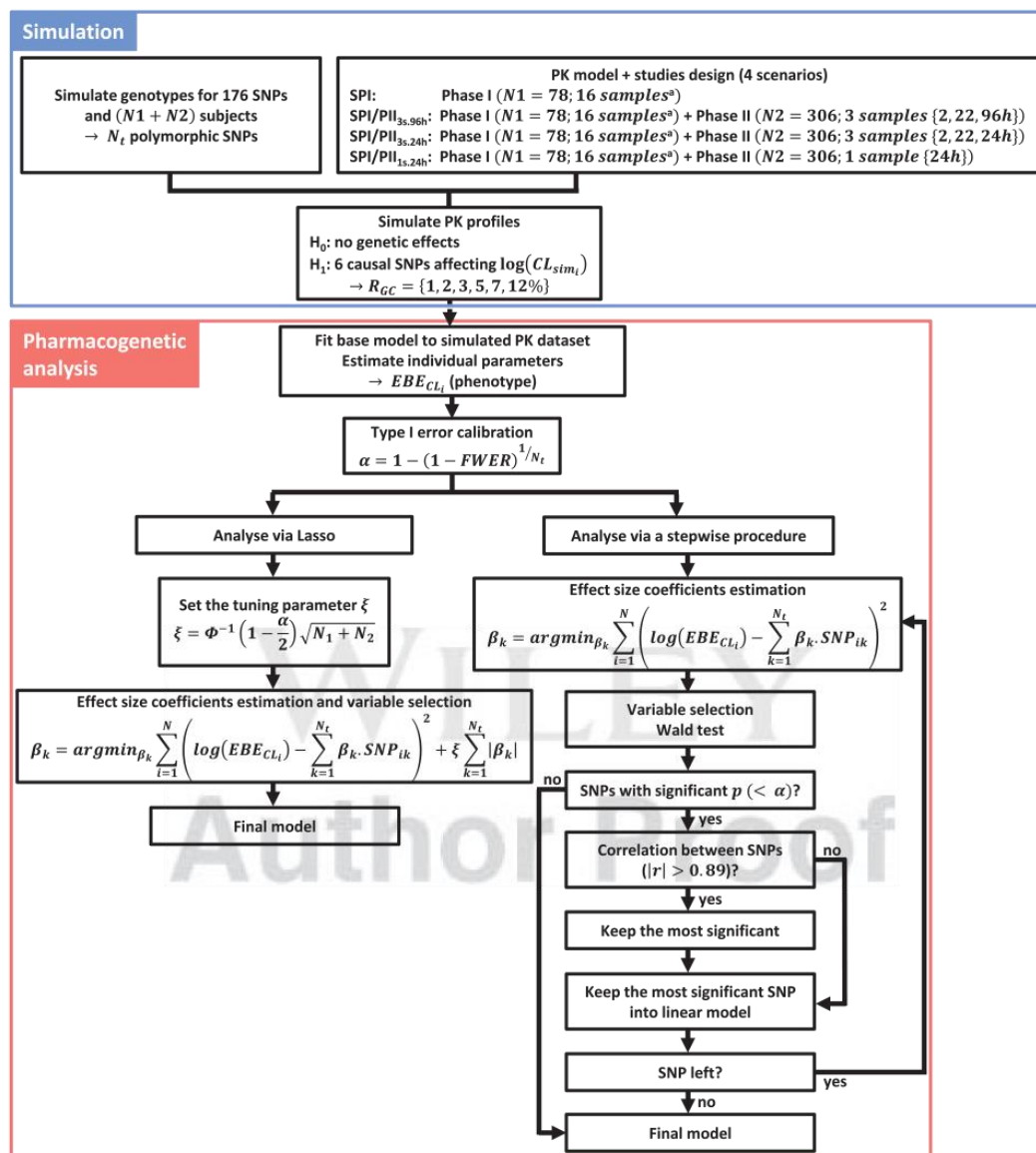
We simulated a phase I study corresponding to the motivating example, including 78 subjects (N1) receiving eight different single doses (5, 10, 20, 50, 100, 200, 400, or 800 units, for respectively 6, 6, 24, 12, 12, 6, 6, and 6 subjects per dose) and sampled 16 times. Three designs of phase II study were simulated. They included 306 subjects (N2), receiving three doses (20, 50, or 100 units, 102 subjects per dose), sampled at steady state. Two phase II studies included three samples per subject, optimized using the PFIM software<sup>18</sup> to ensure a reasonable precision of CL estimates. The last sampling time was limited to 24 hours in one, while a late sample was allowed after the last dose administration in the other. The third study included only one trough concentration (24 hours). We considered four analysis scenarios (**Figure 1**), three combining the phase I and one of the phase II study (respectively, SPI/II<sub>3s,24h</sub>, SPI/II<sub>3s,96h</sub>, and SPI/II<sub>1s,24h</sub>), and one, for comparison, with only the phase I subjects (SPI).

We also investigated the impact of a higher variability on phenotype on the results. For this, we simulated the same four scenarios increasing the IIV on CL to 60% (instead of 25% in previous settings).

### Evaluation

For each analysis scenario, 200 datasets were simulated under  $H_0$  and  $H_1$ .

The ability of the designs to estimate the population and individual parameters under  $H_0$  was first evaluated through estimation bias and  $\eta$ -shrinkage (see details in **Supplementary Material, Supplementary Figures S5–S6**).



**Figure 1** Workflow of the simulation study divided in the simulation (blue box) and analysis part (red box). (a) At 0.5, 1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 48, 72, 96, 120, and 192 hours.  $CL_{sim_i}$ : simulated individual clearance ( $i=1, \dots, N=N_1+N_2$ );  $EBE_{CL_i}$ : empirical Bayes estimate of clearance;  $H_0$ : null scenarios;  $H_1$ : alternative scenarios;  $FWER$ : family-wise error rate;  $N_1$ : number of subjects from the phase I study;  $N_2$ : number of subjects from the phase II study;  $N_t$ : number of polymorphic SNP to analyze;  $p$ :  $P$  value;  $r$ : correlation coefficient between variants;  $R_{GC}$ : genetic component of the interindividual variability;  $SNP_{ik}$ : single nucleotide polymorphism ( $k=1, \dots, N_t$ );  $\alpha$ : type I error per test;  $\beta_k$ : effect size coefficient;  $\xi$ : Lasso tuning parameter.

Under  $H_1$  we evaluated the performance of each scenario in terms of true and false positives counts (TP and FP) and rates (TPR, the proportion of TP detected among the causal variants; and FPR, the proportion of FP detected among all potential false associations) for parameter CL, as well as the probability to detect genetic variants. Assuming that SNPs located on genes coded for metabolism enzymes and transporters affect mostly the drug distribu-

tion and elimination, we also applied association tests on Q, the intercompartmental clearance, and V2, the peripheral volume, separately. Any variants associated to Q and V2 were counted as false positives. The central volume V1 was not considered because it had no random effects.

We also evaluated the loss of genetic signal between simulated and estimated individual clearances, comparing slopes  $b$  of univariate linear regressions on  $\log(CL_{sim_i})$  or



**Table 1** Population values ( $\mu$ ) and interindividual variability ( $\omega$ ) for the model parameters of drug S used in the simulation study

Parameters		$\mu$	$\omega$ (%)
$F^a$	$lmax_F$	0.8	32.9
	$D50_F$	41.7	
$FRAC^b$	$Emax_{FRAC}$	0.45	–
	$D50_{FRAC}$	18.6	
$Tlag_1$		0.401	35.1
$Tk0$		1.59	31.6
$Tlag_2$		22.7	–
$Ka$		0.203	–
$V1$		1520	–
$Q$		147	89.9
$V2$		2130	44.2
$CL$		94.9	25.1
$\sigma_{slope}$ (%)		20	–

<sup>a</sup>For doses < 20 units  $F=1$ , for doses  $\geq 20$  units  $F=1 - \frac{lmax_F (dose - 20)}{D50_F + dose - 20}$ , where dose is the amount administered.

<sup>b</sup> $FRAC = \frac{Emax_{FRAC} \cdot dose}{D50_{FRAC} + dose}$

$F$ : bioavailability;  $FRAC$ : fraction of dose;  $Tk0$ : zero-order absorption duration;  $Tlag_1$ : lag time of zero order absorption;  $Ka$ : first order absorption constant rate;  $Tlag_2$ : lag time of first order absorption;  $V1$ : central compartment volume;  $V2$ : peripheral compartment volume;  $Q$ : intercompartmental clearance;  $CL$ : linear elimination clearance.

$\log(EBE_{CL_i})$  for each causal variant. A relative deviation of the genetic signal  $RD_{signal}$  was computed as follows:

$$RD_{signal}(\%) = \frac{b_{EBE_{CL_i}} - b_{CL_{sim_i}}}{b_{CL_{sim_i}}} \times 100 \quad (3)$$

$RD_{signal}$  quantifies the departure of the estimated genetic signal ( $b_{EBE_{CL_i}}$ ) from the one simulated ( $b_{CL_{sim_i}}$ , see details in **Supplementary Material**).

## RESULTS

### Control of FWER under $H_0$

The Lasso and stepwise procedures both tended to be too conservative, as the FWER was lower than expected in some scenarios (**Table 2**). After an empirical correction by increasing the type I error per SNP  $\alpha$  FWER was properly controlled around 20%. This correction was applied in the corresponding simulation under  $H_1$ . Previous simulations suggested that this decrease in FWER is influenced by correlations between polymorphisms.<sup>6</sup>

**Table 2** Empirical estimates of family-wise error rate under  $H_0$  for both association tests

Method		FWER (%)			
		SPI	SPI/II <sub>3s,96h</sub>	SPI/II <sub>3s,24h</sub>	SPI/II <sub>1s,24h</sub>
Lasso	Without correction <sup>a</sup>	14	17.5	21.5	13.5
Stepwise procedure	Without correction <sup>a</sup>	20	18.5	22.5	15.5
Lasso	After empirical correction <sup>b</sup>	20	19.5	21.5	19.5
Stepwise procedure	After empirical correction <sup>b</sup>	20	20.5	22.5	20.5

<sup>a</sup>Set of empirical family wise error rates (FWER) obtained without correction.

<sup>b</sup>Set of empirical FWER obtained after correction of type I error per tests.

The 95% prediction interval around 20 for 200 simulated datasets is [14.5–25.5].

### Detection of genetic effects

Under  $H_1$  the TPR (**Figure 2**, top left) was higher in scenarios including phase II data (from 22 to 32%) compared to scenario with only phase I data (SPI, 4%) and was the highest in scenario SPI/II<sub>3s,96h</sub>. The FPR was lowest (0.2%) in scenario SPI, where a limited number of SNPs was selected, and only slightly higher in scenarios including phase II data, ranging from 0.6 to 0.8% for both methods. Very few TP were effectively detected in scenario SPI (around 44 for both methods) where the number of subjects was limited ( $N1 = 78$ ) (**Supplementary Tables S2–S3**). By adding more subjects ( $N2 = 306$ ) to the analysis, the number of TP increased sharply. Scenario SPI/II<sub>3s,96h</sub> allowed detecting the largest number of TP (380 or more), while in SPI/II<sub>3s,24h</sub> around 326 TP were detected. In SPI/II<sub>1s,24h</sub> the number of TP was lower (around 270 TP), but remained much higher than scenario SPI with only phase I data. In the same way, the number of FP increased when including phase II data in the analysis, but to a much lesser extent.

With only phase I data, the probability to detect at least one genetic variant on CL was low (**Figure 2**, bottom left), around 20% (SPI). This probability decreased quickly when trying to detect more polymorphisms and reached 0 for 3 variants or more. Adding phase II data to the analysis increased the probability to detect at least one variant about 85% in scenario SPI/II<sub>1s,24h</sub>, and up to 95% in scenario SPI/II<sub>3s,96h</sub>. Scenarios including phase II data showed good detection of one to three SNPs and SPI/II<sub>3s,96h</sub> had always the higher detection. This shows that the major determinant of power is the number of subjects, and that optimizing the design for more informativeness can bring a smaller further improvement. The low probability to detect four SNPs or more ( $\leq 4\%$ ) in scenarios combining phase I and phase II data can be explained by those variants having a very weak impact; polymorphisms only explaining 1, 2, or 3% of the variability of CL.

In **Supplementary Table S4**, the TPR was computed separately for each causal SNP. The variants associated with the lowest  $R_{GC}$  had low TPR, close to the FPR. Thus, the signal associated with these variants was close to the noise created by the noncausal variants.

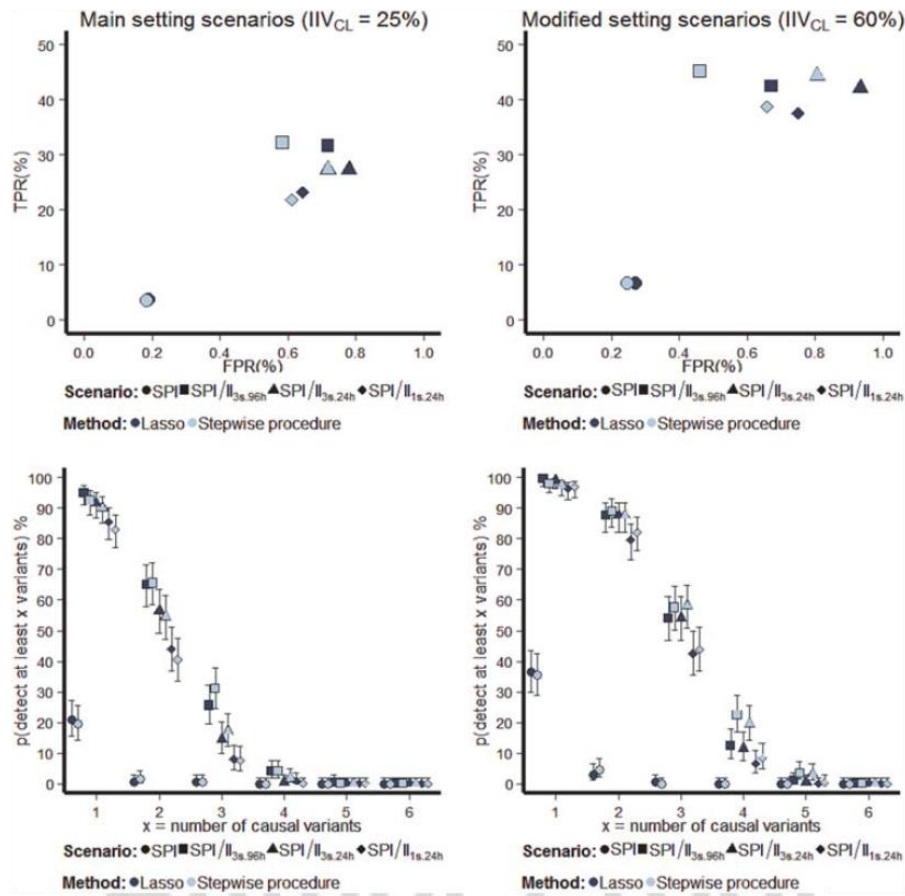
### Shrinkage

Two  $\eta$ -shrinkage estimates were computed using a metric proposed by Bertrand *et al.*<sup>4</sup> based on estimated variances, with respect to the estimate of  $\hat{\omega}^2$  in the dataset; one over the  $\hat{\eta}_i$  from phase I subjects and one over the  $\hat{\eta}_i$  from phase II subjects (**Figure 3**). The  $\eta$ -shrinkage for phase I

T2

F2

F3



**Figure 2** True positive rate (TPR) vs. false positive rate (FPR) under  $H_1$  (top) and probability estimates (points) and 95% confidence interval (bars) to detect at least  $x$  variants explaining the interindividual variability of CL ( $x=1, \dots, 6$ ) under  $H_1$  (bottom) for main scenarios simulated with  $IIV_{CL} = 25\%$  (left) or modified scenarios simulated with  $IIV_{CL} = 60\%$  (right). Different symbols are used for each scenario, and colors denote the Lasso (gray) and the stepwise procedure (light blue).

subjects was low (median = 23%) thanks to the large number of observations per subject. A large range of  $\eta$ -shrinkage estimates for phase II data was observed across analysis scenarios, but was below 50% in scenario SPI/II<sub>3s,96h</sub>.

#### Loss of signal

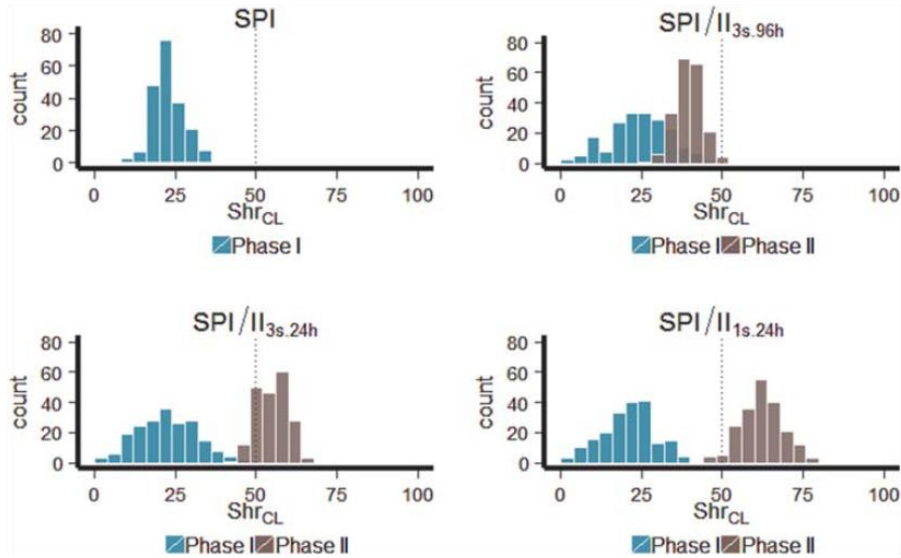
$RD_{signal}$  was always negative for the six SNPs, indicating that part of the signal was lost during the estimation step (Figure 4, top). This loss was smaller in the scenario with phase I data alone (SPI) than in scenarios combining phase I and phase II data. In each scenario the signal loss was of the same magnitude for the six SNPs, regardless of the value of associated  $R_{GC}$ . For phase I data (Figure 4, bottom), the loss was of a constant magnitude across scenarios (median = -30%). For phase II data, in the most informative scenario (SPI/II<sub>3s,96h</sub>), the loss was of a similar magnitude (median = -41%) than the loss in phase I data, where subjects were extensively sampled. The loss was higher in scenario SPI/II<sub>3s,24h</sub> (median = -56%), and even

more when only one time was sampled (SPI/II<sub>1s,24h</sub>, median = -70%).

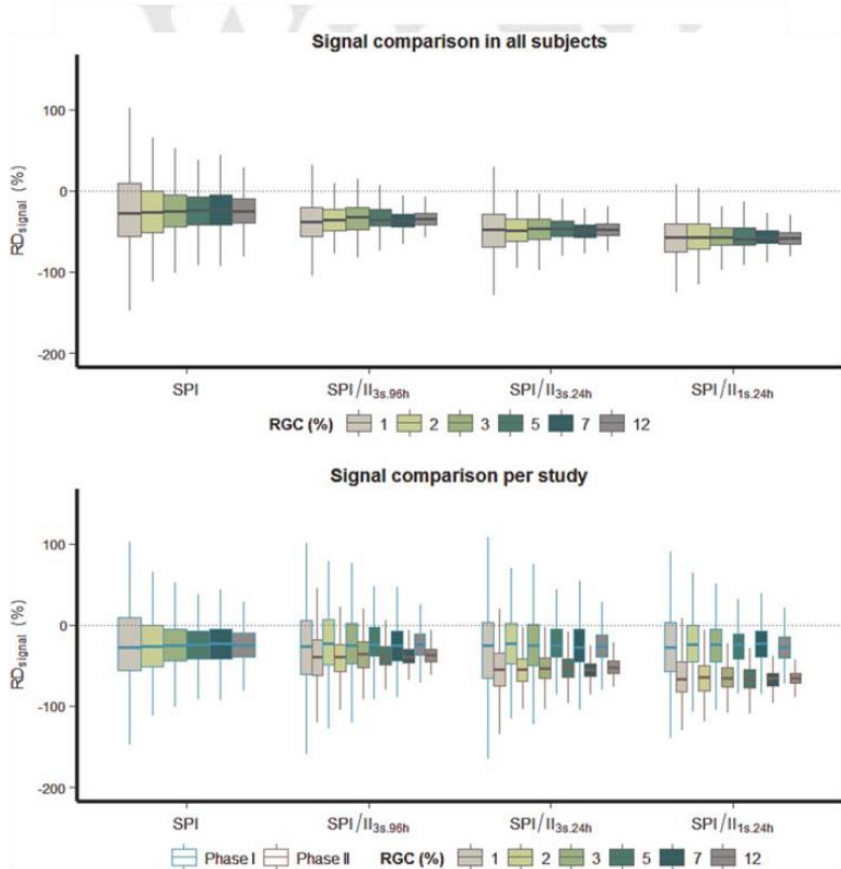
The signal loss and  $\eta$ -shrinkage values changed accordingly across phase II scenarios, while the probability of detection changed in the opposite direction.

#### Influence of the phenotype variance

Increasing IIV for the CL parameter to 60% led to a sharp increase in the number of TP (Supplementary Tables S5–S6), resulting in higher TPR and higher probabilities to detect the causal variants (Figure 2, right), compared to when individual CLs were simulated with a moderate IIV. This higher number of TP is explained first and foremost by the increase in simulated effect sizes, which depended on the variance of interindividual random effects on CL due to nongenetic sources (Equation 2). A second consequence of the larger IIV was that the estimated  $\eta$ -shrinkages became much smaller. Lower  $\eta$ -shrinkages resulted in lower signal losses in all scenarios for phase I and phase II data (Supplementary Figures S7–S8), which again favored a higher probability to detect the genetic effects.



**Figure 3** Distribution of the  $\eta$ -shrinkages on clearance for subjects in the phase I dataset (blue) and for subjects in the phase II dataset (brown), for each main scenario simulated under  $H_0$  with  $IIV_{CL} = 25\%$ .



**Figure 4** Boxplots showing the loss of the signal for genetic effect in the overall population (top), as well as separately for the phase I data (blue borders) and for the phase II data (brown borders) (bottom). A boxplot is shown separately for each main scenario simulated under  $H_1$  with  $IIV_{CL} = 25\%$  as a function of increasing  $R_{GC}$  (boxplots color).



### DISCUSSION

In this work we show and evaluate practical designs to combine data from studies occurring in phase I and II of a drug development. We assess through a simulation study, inspired by a real example, the probability to detect genetic variants and the influence of the phase II study design. We considered phenotypes estimated by NLMEM, which can handle the analysis of heterogeneous data involving sparsely sampled subjects.

Genetic variants are unbalanced and so the amount of information they provide is directly related to the variant allele frequency and the study sample size. On the other hand, PK information depends also on the number and times of sampling which drives the precision of the PK model parameter estimates. A limited number of samples, as in phase II studies, may lead to missing a true association when EBE are used as phenotypes.<sup>9</sup> Savic and Karlsson suggested a more extensive use of the likelihood ratio test (LRT) for covariate selection when  $\eta$ -shrinkage is large, but Combes *et al.* showed that the power to detect a covariate effect is the same with an LRT or a simple correlation test on EBE.<sup>19</sup>

The effect of sample size can be distinctly observed in our simulations. In the context of phase I studies, where the number of subjects is limited, the probability to detect the genetic effects was low, in line with our previous results.<sup>6</sup> The combined analysis of phase I and phase II data allowed a marked improvement in this detection probability, irrespective of the phase II study design. By modifying the design of the phase II data, we highlighted a direct link between  $\eta$ -shrinkage, loss of genetic signal, and probability to detect genetic variants. Our results showed that poor PK information due to the phase II study design results in higher  $\eta$ -shrinkage, which increases the loss of genetic signal at the estimation step and translates to a lower probability to detect genetic variants. The dilution of the individual information by adding subjects with sparse designs to subjects with rich designs increases, as expected, the loss of genetic signal. But this is accompanied with a sharp increase in detection power thanks to a larger sample size.  $\eta$ -shrinkage may also modify the EBE–EBE relationship, falsely inducing or masking correlations between model parameters.<sup>9</sup> This could result in an increased number of false positives associated with other parameters than CL, although in our simulations the number of FP on CL, V2, and Q remained of a similar magnitude across scenarios (Supplementary Table S3), showing no systematic effect.

We assume homogeneity of the PK between subjects simulated for the phase I and the phase II study. In practice, healthy volunteers are often included in phase I, while phase II studies focus on patients. A difference in typical values, for example of CL, between the two populations should not impact the detection power by combination of data, as the association tests use the phenotype variance, provided that the genetic effect is the same and that the model accounts for the systematic difference between clearances. It is more difficult to predict what would happen if the variability of clearance is different in the two popula-

tions, as the magnitude of the shrinkage in each subpopulation could affect the signal detection. When the assumption that the two populations are similar breaks down, we would suggest instead to combine rich and sparse data within the phase II study. Pharmacogenetic studies including a large number of subjects combining sparse and rich designs have already been published,<sup>20,21</sup> showing that the combination of different sampling designs is feasible within the same study to assure more homogeneity.

Situations where pharmacogenetic analyses in PK studies are recommended are described by health authorities.<sup>22</sup> In our work, we simulated a blinded pharmacogenetic analysis, exploring a large number of genetic markers. In real applications, other considerations than the statistical significance of genetic variants such as their physiological and clinical relevance could be factored in the analysis and its interpretation. Lehr *et al.*<sup>23</sup> proposed in their stepwise procedure to select only significant polymorphisms having a physiologic relevance in the final model, and the same constraint could be integrated in penalized regression approaches. The probability to detect genetic variants could also be increased through the targeted inclusion of subjects for a few polymorphisms of interest, but this approach requires hypotheses on which polymorphisms to test, with a risk to miss important associations. We focused in this work on PK variability, which is a part of the variability in drug response. But the conclusions from the simulation study could be extended to pharmacodynamics. A previous survey indicated that most pharmacogenetic analyses in clinical PK studies used a phenotype estimated by NCA and furthermore included a limited number of subjects (fewer than 50 subjects in two-thirds).<sup>6</sup> Authorities in fact recommend studying pharmacogenetics in phase I,<sup>22</sup> where the number of subjects is limited. Our work shows that such analyses do not have the power to detect polymorphisms efficiently, but can generate hypotheses to assess in later studies. A recent simulation work<sup>24</sup> studied the sample size required to detect a binary covariate. They concluded that around 60 subjects combining rich or sparse designs was sufficient to detect the covariate with at least 80% power. Again, our simulations showed that genetic covariates require higher sample sizes because they are highly unbalanced.

In the first series of simulations a moderate IIV on CL (25%) was used, resulting in a low impact of the genetics on PK, since overall 30% of the moderate CL variability was explained by genetic variants. This setting represented a realistic case to challenge the detection of genetic variants through modeling. We also evaluated the same scenarios with a higher IIV for CL, set to 60%. The  $\eta$ -shrinkage was much lower, as a higher IIV downweights the population prior in the combined criterion used to compute EBE. This decrease in CL  $\eta$ -shrinkages resulted in lower signal loss because of the direct relationship between the two. Associated with larger simulated effect sizes, the number of TP and the probability to detect genetic variants increased in these scenarios. The effect of  $\eta$ -shrinkage on the probability to detect genetic effects in these simulations was higher than the one we observed with the main settings,

because the decrease of  $\eta$ -shrinkage was associated with a sharp increase in the number of TP. This shows that our conclusions do not depend on the level of IIV.

This simulation study also confirms the results of our previous work concerning the relative performance of the different association methods.<sup>6</sup> The penalized regression method Lasso and the stepwise procedure showed a similar probability to detect genetic variants in all scenarios. However, Lasso is a slightly more complex method that requires computing the penalty in a first step before testing the associations. In this work we assessed methods to detect genetic effects on EBE, after an initial fit. An algorithm proposed by Lehr *et al.*<sup>23</sup> used univariate regressions to select variants to test in the PK model through LRT. This approach is easy to implement but runtimes depend on the number of iterations leading to the full covariates model. An alternative is to use an integrated approach where effect sizes are estimated and significant variants selected using a penalized regression in the same step<sup>17</sup>; this showed similar performance as the stepwise procedure proposed by Lehr *et al.*, but with longer computing times.<sup>17</sup> The results for the two other penalized regression methods tested in the previous work, ridge regression and Hyper-Lasso, were similar (**Supplementary Tables S7–S9, Figure S9**). None of the methods detected the six SNPs simultaneously, as three of the polymorphisms only explained 1 to 3% of the clearance variability, making them difficult to detect. Because association methods relate the polymorphisms to the phenotype variance, we fixed the variance explained by the causal variants (through the parameter  $R_{GC}$ ) and computed the effect sizes as a function of their allelic frequencies. For a given  $R_{GC}$  an infrequent polymorphism was therefore associated with higher effect sizes. This reflects that a clinically relevant polymorphism (with a high impact on PK), present in few subjects because of its low frequency, will explain a limited proportion of the phenotype variance. Detecting such polymorphisms is crucial to identify subpopulations at risk but require much larger sample sizes, as in genome-wide studies.<sup>25</sup> As an example, the rs3918290 polymorphism from gene *DPYD* has a frequency lower than 1%, but results in a deficient dihydropyrimidine dehydrogenase activity associated with a 40% decrease of the maximum conversion capacity of the chemotherapeutic drug 5-fluorouracil,<sup>26</sup> resulting in severe toxicities.

The power to detect polymorphisms is also closely related to the type I error chosen for the analysis. In a context of exploratory analyses, we fixed the global type I error to 20%. But using the Sidak correction the significance thresholds were finally lower than 0.1% for each test, so that only strong effects of causal variants will be detected, and our simulations show that polymorphisms explaining a limited part of the phenotype variance are not detected. Approaches based on FWER and corrections such as Bonferroni or Sidak are easy to implement but are conservative and may reduce the power of analyses, but limit the number of polymorphisms to test in later confirmatory trials. In practice, other corrections for type I error could be considered, as permutation methods that are more time-consuming but less conservative.

Although this correction was conservative and was calibrated under  $H_0$  to control the FWER, the proportion of FP under  $H_1$  among selected variants was higher than the expected 20%. This could reflect the correlations between polymorphisms we simulated.

To make more specific recommendations for study designs is difficult because it is closely related to the developed drug. In our simulations a late sample allowed larger information on the elimination phase to estimate CL. This result can be generalized to pharmacogenetic studies involving clearance and drugs with a long half-life. Taking a late sample requires suspending treatment long enough to observe a decrease in concentrations, which may not be possible in patients from phase II trials.

In any case, it is essential that the sampling protocol, although limited, is as informative as possible to minimize the estimation error and shrinkage in individual parameters estimation. The detection of genetic polymorphisms could highly benefit from the use of larger sample sizes through combined analysis and optimized design.<sup>18,27</sup>

In conclusion, this work confirmed the very limited likelihood that weak genetic effects can be detected in a typical phase I study, due to the small sample size. Such studies have to be considered only as hypothesis-generating.<sup>28</sup> On the basis of our results in terms of detection probability when analyzing together data from phase I and phase II studies, we claim that phase II is the best moment to identify the impact of genetic variants on drug response. It would be less efficient to start the study of the pharmacogenetics of a new drug in phase III trials or in postmarketing, because these take place too late in drug development<sup>29</sup> and the new treatment could be administered in nonresponders or expose subjects to high toxicities. Furthermore, genetic subpopulations can be better targeted and potentially some subjects excluded from the study to increase the efficacy and reduce the risk of toxicity of the drug in these phase III studies.

**Acknowledgments.** Adrien Tessier received funding from Institut de Recherches Internationales Servier. The authors thank Laurent Ripoll and Bernard Walther from Institut de Recherches Internationales Servier for their advice in pharmacogenetics. The authors also thank Hervé Le Nagard for the use of the computer cluster services hosted on the “Centre de Biomodélisation UMR1137.” This work received the Lewis Sheiner Student Award from the Population Approach Group in Europe (PAGE) committee and was presented as oral communication in the Lewis Sheiner Student Session at the 24th annual PAGE meeting: Tessier, A. *et al.* Modelling pharmacogenetic data in population studies during drug development. *PAGE Abstract* #3333 <<http://www.page-meeting.org/?abstract=3333>> (2015).

**Conflict of Interest.** Adrien Tessier has a research grant from Institut de Recherches Internationales Servier and the French government. Marylore Chenel works for Institut de Recherches Internationales Servier, heading the department of Clinical Pharmacokinetics and Pharmacometrics.

**Author Contributions.** A.T., J.B., M.C., and E.C. wrote the article; A.T., J.B., M.C., and E.C. designed the research; A.T., J.B., M.C., and E.C. performed the research; A.T., J.B., M.C., and E.C. analyzed the data.



- Aarons, L. Population pharmacokinetics: theory and practice. *Br. J. Clin. Pharmacol.* **32**, 669–670 (1991).
- Motulsky, A.G. Drugs and genes. *Ann. Intern. Med.* **70**, 1269–1272 (1969).
- Guo, Y., Shafer, S., Weller, P., Usuka, J. & Peltz, G. Pharmacogenomics and drug development. *Pharmacogenomics* **6**, 857–864 (2005).
- Bertrand, J., Comets, E., Laffont, C.M., Chenel, M. & Mentré, F. Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm. *J. Pharmacokinet. Pharmacodyn.* **36**, 317–339 (2009).
- Bertrand, J., Comets, E., Chenel, M. & Mentré, F. Some alternatives to asymptotic tests for the analysis of pharmacogenetic data using nonlinear mixed effects models. *Biometrics* **68**, 146–155 (2012).
- Tessier, A., Bertrand, J., Chenel, M. & Comets, E. Comparison of nonlinear mixed effects models and noncompartmental approaches in detecting pharmacogenetic covariates. *AAPS J.* **17**, 597–608 (2015).
- Sheiner, L.B., Rosenberg, B. & Melmon, K.L. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput. Biomed. Res. Int. J.* **5**, 411–459 (1972).
- Rowland, M. & Tozer, T.N. Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications. (Wolters Kluwer Health/Lippincott William & Wilkins, Philadelphia, 2011).
- Savic, R.M. & Karlsson, M.O. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *AAPS J.* **11**, 558–569 (2009).
- Mandema, J.W., Verotta, D. & Sheiner, L.B. Building population pharmacokinetic—pharmacodynamic models. I. Models for covariate effects. *J. Pharmacokinet. Biopharm.* **20**, 511–528 (1992).
- Tessier, A., Bertrand, J., Fouliard, S., Comets, E. & Chenel, M. High-throughput genetic screening and pharmacokinetic population modeling in drug development. (2013). Abstract #2836 at <www.page-meeting.org/?abstract=2836.>
- International HapMap Consortium The International HapMap Project. *Nature* **426**, 789–796 (2003).
- Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).
- Bertrand, J. & Balding, D.J. Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models. *Pharmacogenet. Genomics* **23**, 167–174 (2013).
- Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1994).
- Hoggart, C.J., Whittaker, J.C., De Iorio, M. & Balding, D.J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
- Bertrand, J., De Iorio, M. & Balding, D.J. Integrating dynamic mixed-effect modelling and penalized regression to explore genetic association with pharmacokinetics. *Pharmacogenet. Genomics* **25**, 231–238 (2015).
- Bazzoli, C., Retout, S. & Mentré, F. Design evaluation and optimisation in multiple response nonlinear mixed effect models: PFIM 3.0. *Comput. Methods Programs Biomed.* **98**, 55–65 (2010).
- Combes, F., Retout, S., Frey, N. & Mentré, F. Powers of the likelihood ratio test and the correlation test using empirical Bayes estimates for various shrinkages in population pharmacokinetics. *CPT Pharmacomet. Syst. Pharmacol.* **3**, 1–9 (2014).
- Chou, M. *et al.* Population pharmacokinetic-pharmacogenetic study of nevirapine in HIV-infected Cambodian patients. *Antimicrob. Agents Chemother.* **54**, 4432–4439 (2010).
- Bertrand, J. *et al.* Multiple genetic variants predict steady-state nevirapine clearance in HIV-infected Cambodians. *Pharmacogenet. Genomics* **22**, 868–876 (2012).
- EMA Guideline on the Use of Pharmacogenetic Methodologies in the Pharmacokinetic Evaluation of Medicinal Products. (2012).
- Lehr, T., Schaefer, H.-G. & Staab, A. Integration of high-throughput genotyping data into pharmacometric analyses using nonlinear mixed effects modeling. *Pharmacogenet. Genomics* **20**, 442–450 (2010).
- Klopprogge, F., Simpson, J.A., Day, N.P.J., White, N.J. & Tarning, J. Statistical power calculations for mixed pharmacokinetic study designs using a population approach. *AAPS J.* **16**, 1110–1118 (2014).
- Takeuchi, F. *et al.* A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet.* **5**, e1000433 (2009).
- Kuilenburg, A. B. P. van *et al.* Evaluation of 5-fluorouracil pharmacokinetics in cancer patients with a c.1905 + 1G>A mutation in DPYD by means of a Bayesian limited sampling strategy. *Clin. Pharmacokinet.* **51**, 163–174 (2012).
- Combes, F.P., Retout, S., Frey, N. & Mentré, F. Prediction of shrinkage of individual parameters using the bayesian information matrix in non-linear mixed effect models with evaluation in pharmacokinetics. *Pharm. Res.* **30**, 2355–2367 (2013).
- Bromley, C.M. *et al.* Designing pharmacogenetic projects in industry: practical design perspectives from the Industry Pharmacogenomics Working Group. *Pharmacogenomics J.* **9**, 14–22 (2009).
- O'Donnell, P.H. & Stadler, W.M. Pharmacogenomics in early-phase oncology clinical trials: is there a sweet spot in phase II? *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **18**, 2809–2816 (2012).

© 2015 The Authors CPT: Pharmacometrics & Systems Pharmacology published by Wiley Periodicals, Inc. on behalf of American Society for Clinical Pharmacology and Therapeutics. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Supplementary information accompanies this paper on the CPT: Pharmacometrics & Systems Pharmacology website (<http://www.wileyonlinelibrary.com/psp4>)

### **5.3. Matériel supplémentaire**

**Supplementary material to ‘Combined analysis of phase I and phase II data to enhance the power of pharmacogenetic tests’ by Tessier et al.**

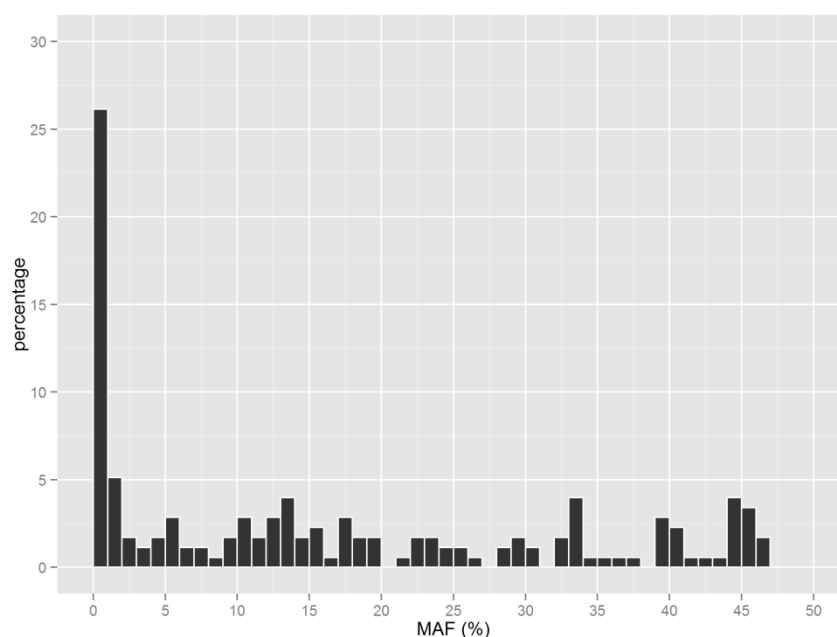
**The supplementary material for the paper covers the following elements:**

- Additional details on the simulation study materials and methods
  - Simulated polymorphisms
  - Pharmacokinetic model used for simulations of concentrations profiles
  - Computation of the correlation coefficient in the stepwise procedure
  - Evaluation
  - Softwares
- Additional results for scenarios presented in the main manuscript
  - Simulated genetic effects
  - Estimation bias
- Additional results for the two other penalised regression methods
- Additional references
- Supplementary figures legends

## Additional details on the simulation study materials and methods

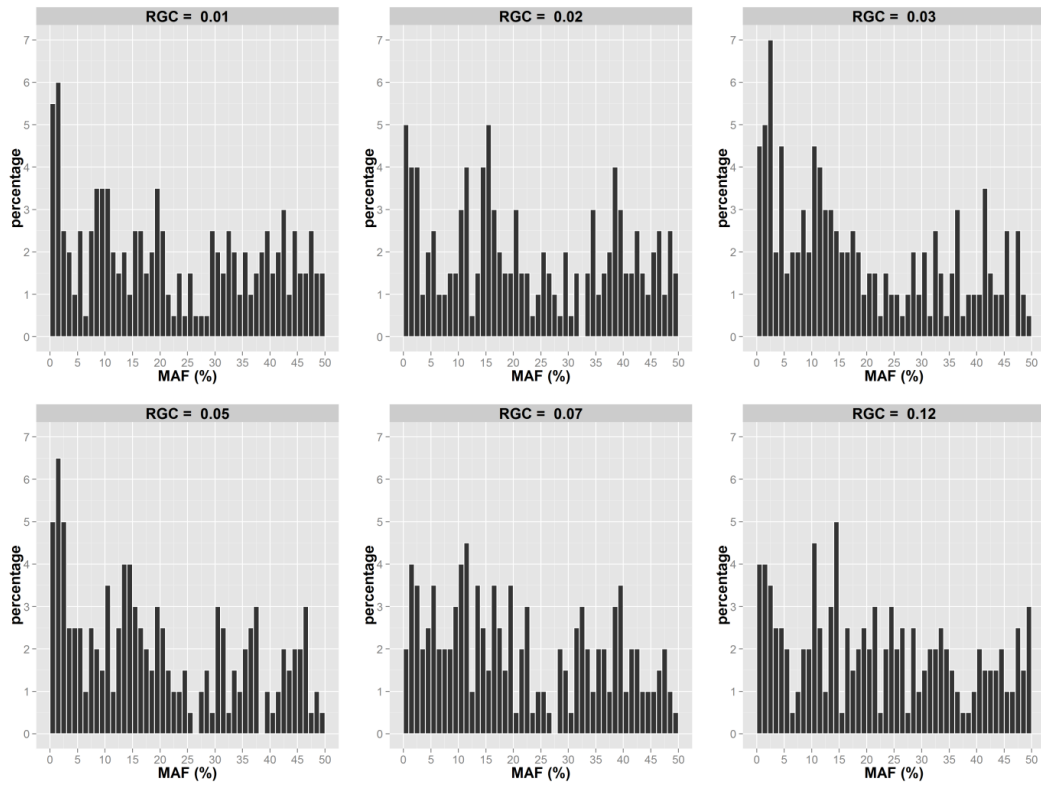
### Simulated polymorphisms

The genotypes for 176 SNPs were simulated using the Hapgen2 software and a reference Hapmap panel (Hapmap 3 release 2) for a Caucasian population. Hapgen2 resamples haplotypes from the reference panel to simulate new haplotypes as an imperfect mosaic of reference haplotypes using a Hidden Markov Model, mimicking the effect of recombination<sup>1</sup>.

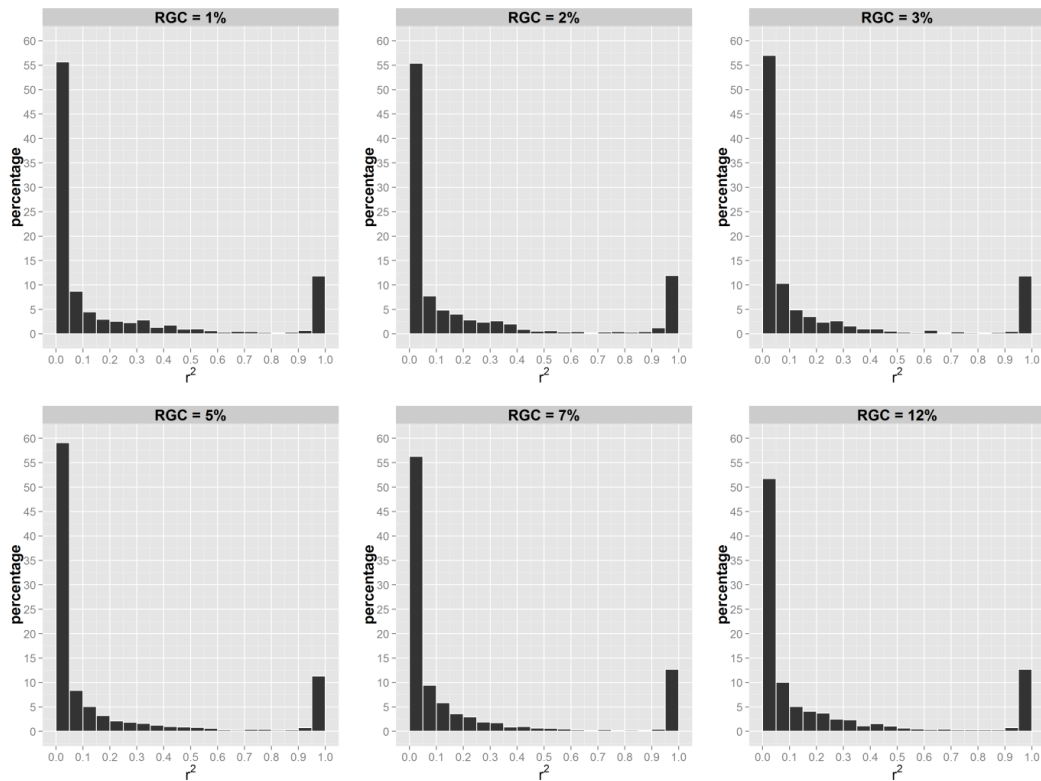


**Figure S1.** Distribution of the mean of Minor Allele frequencies (MAF) for the 176 SNPs simulated across the 200 dataset.





**Figure S2.** Distribution of Minor Allele frequencies (MAF) for the 6 causal SNPs simulated across the 200 dataset as a function of their respective  $R_{GC}$ .

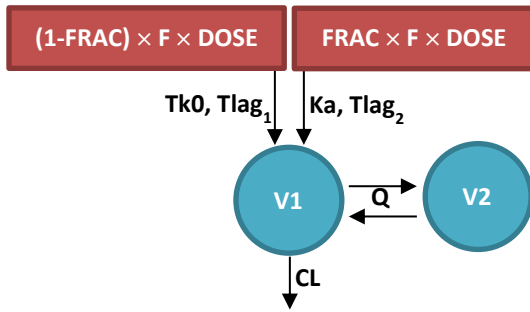


**Figure S3.** Distribution of correlations (absolute value of  $r$ ) between the causal SNP and other SNPs in linkage disequilibrium as a function of the  $R_{GC}$  associated with the causal SNP across the 200 simulated dataset.

Simulated polymorphisms have the same frequencies (**Figure S1- S2**) and correlations patterns than the reference HapMap panel (**Figure S3**). On average, 21% of the correlations between the causal SNP and SNPs in linkage disequilibrium have correlation coefficient  $|r|$  higher than 0.89, regardless to the  $R_{GC}$  associated. This can be attributed to the DNA chip design, where groups of SNPs from the same gene, so physically close, are present.

#### Pharmacokinetic model used for simulations of concentrations profiles

PK profiles were simulated using the model and parameters estimates best describing the nonlinear PK of the motivating data example. The structural model includes two-compartment for disposition, a double absorption function nonlinear with dose and a linear elimination (**Figure S4**).



**Figure S4.** Structural pharmacokinetic model of drug S. Double absorption compartments in red and distribution compartments in blue. F: bioavailability ; FRAC: fraction of dose ; Tk0: zero order absorption constant rate ; Tlag<sub>1</sub>: lag time of zero order absorption ; Ka: first order absorption constant rate ; Tlag<sub>2</sub>: lag time of first order absorption ; V1: central compartment volume ; V2: peripheral compartment volume ; Q: intercompartmental clearance ; CL: linear elimination clearance

The simulated data sets were fitted using NLMEM, as in the original analysis. The drug concentration at time  $t_{ij}$  in the  $i^{\text{th}}$  subject ( $i = 1, \dots, N, j = 1, \dots, n$ ),  $y_{ij}$ , was assumed to follow a nonlinear function  $f$ , such as:

$$y_{ij} = f(\theta_i ; \xi_{ij}) + \varepsilon_{ij} \quad (1)$$

where  $\theta_i$  is the vector of individual parameters and  $\varepsilon_{ij}$  the residual error. We assume that  $\theta_i$  follow a log-normal distribution,  $\theta_i = \mu e^{\eta_i}$ , with  $\mu$  the average population parameter and  $\eta_i$  the random effect in individual  $i$ .  $\xi_{ij}$  is the elementary design for the subject  $i$ , i.e. a vector of  $n$  sampling times ( $\xi_{ij} = t_{i,1}, \dots, t_{i,n}$ ). We typically assume  $\eta_i \sim N(0, \omega^2)$  and  $\varepsilon_{ij} \sim N(0, \sigma^2)$ .

The SAEM estimation algorithm<sup>2</sup> implemented in the Monolix software ([www.lixoft.eu](http://www.lixoft.eu)) was used to estimate the population parameters ( $\mu$ ,  $\omega^2$  and  $\sigma^2$ ) by maximum likelihood. *Maximum a Posteriori* method (MAP) estimates of the individual parameters were obtained by a Bayesian approach<sup>3</sup>. The individual parameters are also called Empirical Bayes Estimates (EBE).

### Computation of the correlation coefficient in the stepwise procedure

The Pearson's correlation coefficient  $r$  is a measure of the linkage disequilibrium between two markers. For example for two SNPs with respective alleles A/a and B/b, the correlation coefficient is computed as follows:

$$r = \frac{p_{AB} - p_A \cdot p_B}{\sqrt{p_A(1 - p_A) \cdot p_B(1 - p_B)}} \quad (2)$$

where  $p_{AB}$  is the frequency of the haplotype AB and  $p_A$  and  $p_B$  the frequencies of the alleles A and B respectively<sup>4</sup>. The coefficient  $r$  can take values between -1 and 1. If this value is close to 0, the SNPs are in equilibrium. On the contrary if  $|r|$  tend to 1, the markers are in strong disequilibrium. The Pearson's correlation coefficient was computed using the PLINK software<sup>5</sup>, which also compute the square of the coefficient, *i.e.* the coefficient of determination, to remove the arbitrary sign introduced.

### Evaluation

For each analysis scenario, 200 data sets were simulated under each hypothesis  $H_0$  and  $H_1$ .

We first evaluated the ability of the designs to estimate the population and individual parameters under  $H_0$  through diagnostic plots: the population estimates of CL ( $\hat{\mu}_{CL_{est}}$ ) were compared to the population CL value ( $\mu_{CL_{sim}}$ ) used in simulations through the relative estimation errors ( $REE_{pop}$ ) estimates obtained on each of the 200 data sets:

$$REE_{pop}(\%) = \frac{\hat{\mu}_{CL_{est}} - \mu_{CL_{sim}}}{\mu_{CL_{sim}}} \times 100 \quad (3)$$

The individual CL simulated ( $CL_{sim_i}$ ) using parameters distributions from the motivating example to compute PK profiles were also compared with CL estimated by NLMEM on the simulated data ( $EBE_{CL_i}$ ). In each data set, we computed the  $REE_{indiv}$  between estimated and simulated individual clearances in the same way, as:

$$REE_{indiv}(\%) = \frac{EBE_{CL_i} - CL_{sim_i}}{CL_{sim_i}} \times 100 \quad (4)$$

To evaluate the shrinkage on random effects ( $Sh_\eta$ ), Bertrand et al<sup>6</sup> used a metric based on estimated variances:

$$Sh_\eta^{var} = 1 - \frac{var(\hat{\eta}_i)}{\hat{\omega}^2} \quad (5)$$

where  $\hat{\eta}_i$  are the EBE and  $\hat{\omega}^2$  the estimate of variance of random effects from the population step. Under  $H_0$ , we computed this metric which reflects the quantity of information brought by each subject to the estimation of individual parameters. When shrinkage is high ( $Sh_\eta^{var} \geq 50\%$ ), the individual estimates are shrunk toward  $\hat{\mu}_{CL_{est}}$ , and covariate tests based on individual parameters may be misleading for instance a covariate relationship may be masked<sup>7</sup>.

The number of true positives ( $TP$ ) was the count of causal variants associated to CL (maximum of 1200 over the 200 simulations) and the number of false positives ( $FP$ ) the count of non-causal variants associated to CL and any variants associated to Q and V2. A 95% confidence interval was estimated assuming the number of TP or FP follows a Poisson distribution. The true positive rate ( $TPR$ ) and the false positive rate ( $FPR$ ) were calculated as follow:

$$TPR = \frac{TP}{TP + FN} \quad (6.1)$$

$$FPR = \frac{FP}{FP + TN} \quad (6.2)$$

where  $FN$  and  $TN$  are respectively the count of false and true negatives. The probability to detect a given number  $x$  ( $x = 1, \dots, 6$ ) of the 6 SNPs which were associated to CL by simulation was estimated by the percentage of data sets where  $x$  or more SNPs were selected under  $H_1$ .

To quantify the loss of genetic signal between simulated and estimated individual clearances, we performed linear univariate regressions on  $CL_{sim_i}$  or  $EBE_{CL_i}$  as a function of the different causal SNPs:

$$\log(CL_{sim_i}) = a + b_{CL_{sim_i}} \cdot SNP_{ik} \quad (7)$$

$$\log(EBE_{CL_i}) = a + b_{EBE_{CL_i}} \cdot SNP_{ik}$$

where  $a$  represents an intercept,  $b_{CL_{sim_i}}$  the regression slope on simulated clearance ( $CL_{sim_i}$ ),  $b_{EBE_{CL_i}}$  the regression slope on estimated clearance ( $EBE_{CL_i}$ ), and  $SNP_{ik}$  the genotype of the  $k^{th}$  causal variant ( $k = 1, \dots, 6$ ) for subject  $i$ ;  $SNP_{ik} = \{0, 1, 2\}$ . The slopes associated to the 6 causal variants were computed separately for each simulated dataset, regardless of their significance with the different association methods. They represent the increase in clearance due to mutated alleles (genetic signal), and were compared to quantify the departure of the estimated genetic signal from the simulated value (relative deviation,  $RD_{signal}$ ):

$$RD_{signal}(\%) = \frac{b_{EBE_{CL_i}} - b_{CL_{sim_i}}}{b_{CL_{sim_i}}} \times 100 \quad (8)$$

### Softwares

The C++ software Hapgen version 2.1.2<sup>8</sup> was used to simulate SNPs from reference haplotype panels provided by the Hapmap Project. Among the 176 SNPs present on the DNA microarray, 55 were also present in the Hapmap reference panel. For the others, we chose the closest variant in the Hapmap database. The SNPs selected from the Hapmap database to target the SNPs from the IRIS DNA microarray were very close to the original ones with a median of 1730 bases pair departure. ([mathgen.stats.ox.ac.uk/genetics\\_software/hapgen/hapgen2.html](http://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html))

The software Monolix version 4.2.2 ([www.lixoft.eu](http://www.lixoft.eu)) implements the SAEM algorithm for parameters estimation in nonlinear mixed effects models. It was used here to derive the EBE using the PK model and the simulated PK profiles.

The C++ software PLINK version 1.07<sup>5</sup> ([pngu.mgh.harvard.edu/~purcell/plink/](http://pngu.mgh.harvard.edu/~purcell/plink/)), is a free and open-source whole genome association analysis toolset, which was used to performed the stepwise procedure.

The statistical software R version 2.15.2<sup>9</sup> ([www.R-project.org/](http://www.R-project.org/)) was used to simulate the PK profiles, as well as for all the statistical analyses and creation of graphical material. Some R

packages were also used such as “ridge”<sup>10</sup>, a package for ridge regression with automatic selection of the regularization parameter.

The C++ program HyperLasso<sup>11</sup> ([www.ebi.ac.uk/projects/BARGEN](http://www.ebi.ac.uk/projects/BARGEN)) was used for the Lasso as well as the HyperLasso regression methods. Of note, the phenotype is standardized within this program so  $\sigma$  is equal to 1 in the formula to set  $\gamma$  for HyperLasso.

## Additional results for scenarios presented in the main manuscript

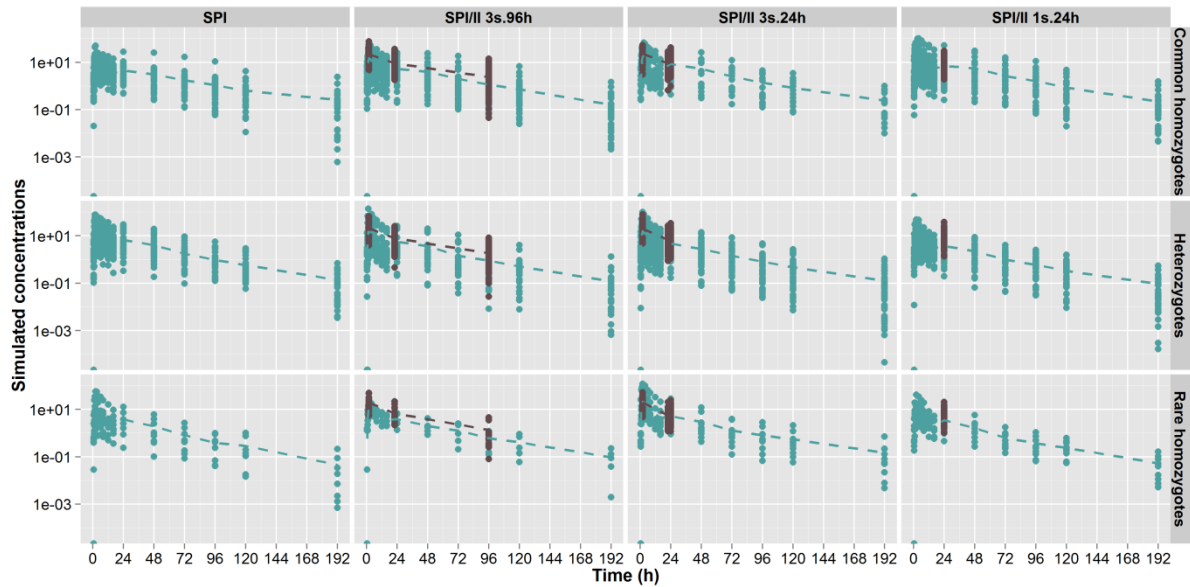
### Simulated genetic effects

**Table S1.** 10<sup>th</sup>, 50<sup>th</sup> and 90<sup>th</sup> percentiles of simulated effect sizes and simulated individual clearances as a function of the genotype of the causal SNP explaining 1 or 12% of clearance variability, for scenarios SPI and SPI/II<sub>3s.96h</sub>.

Scenario	RGC	Percentile	$\beta_k$	Genotype-based typical CL <sup>a</sup>		
				<i>SNP</i> = 0	<i>SNP</i> = 1	<i>SNP</i> = 2
SPI	1%	0.1	0.03	94.9	97.8	100.8
		0.5	0.04	94.9	98.4	102.1
		0.9	0.1	94.9	104.8	115.8
	12%	0.1	0.11	94.9	106	118.3
		0.5	0.15	94.9	109.9	127.4
		0.9	0.35	94.9	135.3	193
SPI/II <sub>3s.96h</sub>	1%	0.1	0.03	94.9	97.8	100.8
		0.5	0.04	94.9	98.6	102.4
		0.9	0.11	94.9	106.1	118.6
	12%	0.1	0.11	94.9	106	118.3
		0.5	0.14	94.9	108.8	124.7
		0.9	0.34	94.9	133.1	186.6

a. the typical clearance was computed as  $CL = \mu_{CL} \cdot e^{\beta_k \cdot SNP_k}$ , where  $\mu_{CL}$  is the typical value of CL in the overall population (94.9),  $\beta_k$  the effect size associated to the causal  $SNP_k$  ( $SNP_k = \{0,1,2\}$ ).

**Figure S5** represents the concentration profiles from one simulated data set under the alternative hypothesis  $H_1$  for all scenarios, sorted by genotypes of the SNP explaining 12% of the clearance interindividual variability. As expected because of the moderate variability for CL (25%), the trough concentrations were slightly lower in heterozygote subjects, and decreased a little further in rare homozygotes, for phase I as well as phase II data in all scenarios. The impact of genetics on PK was difficult to observe based on individual profiles only, justifying the use of sophisticated approaches as NLMEM.



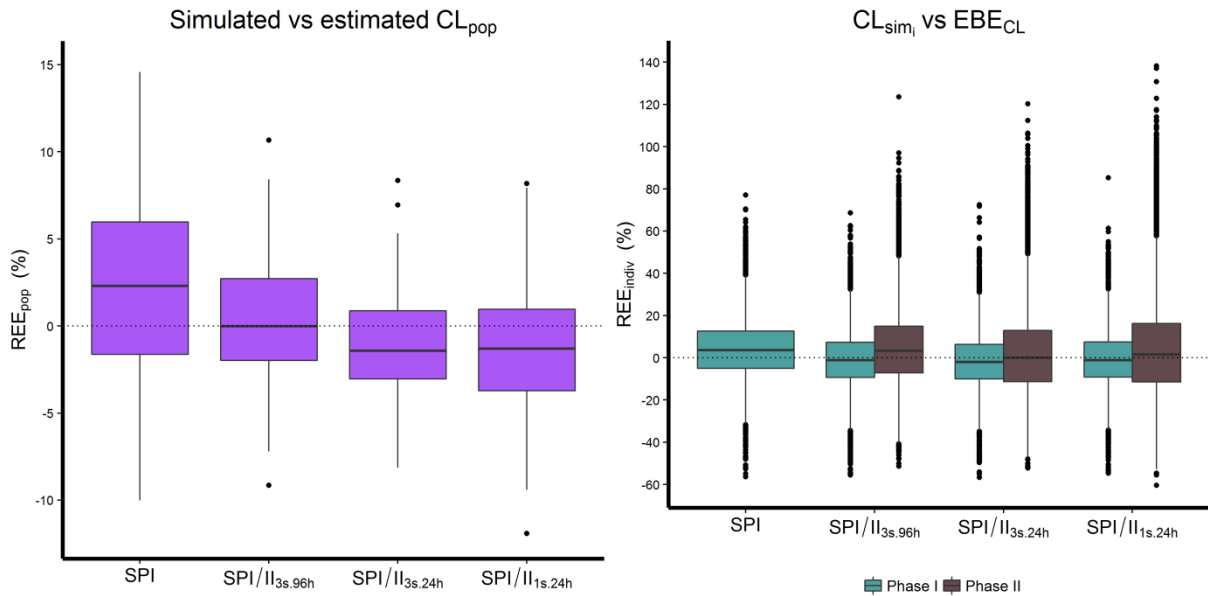
**Figure S5.** Phase I (blue dots) and phase II study (brown dots) concentrations versus time for an example data set for each main scenario simulated with  $IIV_{CL} = 25\%$  under  $H_1$ . Mean profiles (dashed lines) and concentrations are plotted in log-scale on the Y-axis. Concentrations are sorted for the SNP explaining 12% of CL interindividual variability; common homozygotes (*top*), heterozygotes (*middle*) and rare homozygotes (*bottom*).

### Estimation bias

On average, the bias on the population estimate of CL was less than 5% for all scenarios (**Figure S6, left**), showing that this parameter is well estimated in all our settings. Combining phase II data with the phase I data brought a small increase in precision, as shown from the smaller range of REE. This was true even in scenario SPI/II<sub>1s,24h</sub>, where the information on CL from the time point at 24h is very limited, but the best compromise was obtained in the scenario with a late sample at 96h (SPI/II<sub>3s,96h</sub>), where there was no bias on the estimate and the precision was less than 10% for most simulated datasets. Overall, adding phase II data had no major impact on the estimation of population CL.

In contrast to population estimates,  $REE_{indiv}$  were much more widely spread out. However, medians for both phase I and phase II data were close to 0 (**Figure S6, right**). The range of  $REE_{indiv}$  was constant across analysis scenarios. For phase I data it ranged from -55% to 75%, and from -60% to 130% for phase II data, due to the differences in number of sampling times per subject. The bias in mean  $REE_{indiv}$  was smaller in phase II data with three samples (SPI/II<sub>3s,24h</sub> and SPI/II<sub>3s,96h</sub>) than in phase II data with one sample (SPI/II<sub>1s,24h</sub>).





**Figure S6.** Boxplots of the simulated and estimated population clearance in all scenarios simulated under  $H_0$  ( $REE_{pop}$ , left). Boxplots of the simulated and estimated individual clearance in all scenarios simulated under  $H_0$  ( $REE_{indiv}$ , right) for subjects in the phase I (blue) and the phase II study (brown).

### Detection of genetic effects

**Table S2.** True positives counts under  $H_1$  for Lasso and stepwise procedure methods in all scenarios

Method	TP (CL)			
	SPI	SPI/II <sub>3s.96h</sub>	SPI/II <sub>3s.24h</sub>	SPI/II <sub>1s.24h</sub>
Lasso	44 [32;59]	380 [343;420]	326 [292;363]	277 [245;312]
Stepwise procedure	43 [31;58]	386 [348;426]	327 [293;364]	262 [231;296]

Total number of true positives (TP) with their 95% confidence interval under the alternative hypothesis.

On 200 simulated data sets, overall 1200 SNPs were set to impact clearance (maximum TP number).

**Table S3.** False positives counts on CL, Q and V2 under  $H_1$  for Lasso and stepwise procedure methods in all scenarios

Method	$S_{\text{phaseI}}$			SPI/II <sub>3s.96h</sub>			SPI/II <sub>3s.24h</sub>			SPI/II <sub>1s.24h</sub>		
	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)
Lasso	21 [13;32]	35 [24;49]	8 [3;16]	131 [110;155]	40 [29;54]	73 [57;92]	122 [101;146]	112 [92;135]	31 [21;44]	120 [99;43]	68 [53;86]	31 [21;44]
Stepwise procedure	18 [11;28]	35 [24;49]	9 [4;17]	91 [73;112]	47 [35;63]	61 [47;78]	84 [67;104]	107 [88;129]	53 [40;69]	84 [67;104]	74 [58;93]	50 [37;66]

Total number of false positives (FP) with their 95% confidence interval under the alternative hypothesis.

**Table S4.** True positive rates for each causal variant ( $TPR_{R_{GC}}$ ) as a function of the  $R_{GC}$ , and false positive rates (FPR) under  $H_1$  for Lasso and stepwise procedure methods in all scenarios.

Scenario	Method	FPR	$TPR_{R_{GC}}$					
			$R_{GC} = 1\%$	$R_{GC} = 2\%$	$R_{GC} = 3\%$	$R_{GC} = 5\%$	$R_{GC} = 7\%$	$R_{GC} = 12\%$
SPI	Lasso	0.2	1.5	0.5	1.5	5	4	9.5
SPI	Stepwise procedure	0.2	1.5	0.5	1.5	5	4	9
SPI/PII <sub>3s.96h</sub>	Lasso	0.7	2.5	4	13.5	31	52	87
SPI/PII <sub>3s.96h</sub>	Stepwise procedure	0.6	3.5	6	13	36.5	51.5	82.5
SPI/PII <sub>3s.24h</sub>	Lasso	0.8	3	3.5	8.5	25	41	82
SPI/PII <sub>3s.24h</sub>	Stepwise procedure	0.7	4	4	9	26.5	42	78
SPI/PII <sub>1s.24h</sub>	Lasso	0.6	3.5	5	11	21.5	32.5	65
SPI/PII <sub>1s.24h</sub>	Stepwise procedure	0.6	2.5	4	10	21.5	31.5	61.5

FPR: proportion of false positives detected among all potential false associations in the 200 dataset.

$TPR_{R_{GC}}$ : proportion of true positives detected for a causal variant in the 200 dataset (equivalent to the probability to detect this variant).

### Influence of the phenotype variance

**Table S5.** True positives counts under  $H_1$  for Lasso and stepwise procedure methods in all scenarios simulated with  $IIV_{CL} = 60\%$ .

Method	TP (CL)			
	$SPI_{60\%}$	$SPI/II_{3s.96h\_60\%}$	$SPI/II_{3s.24h\_60\%}$	$SPI/II_{1s.24h\_60\%}$
Lasso	80 [63;100]	509 [466;555]	505 [462;551]	450 [409;494]
Stepwise procedure	80 [63;100]	541 [496;589]	531 [487;578]	464 [423;508]

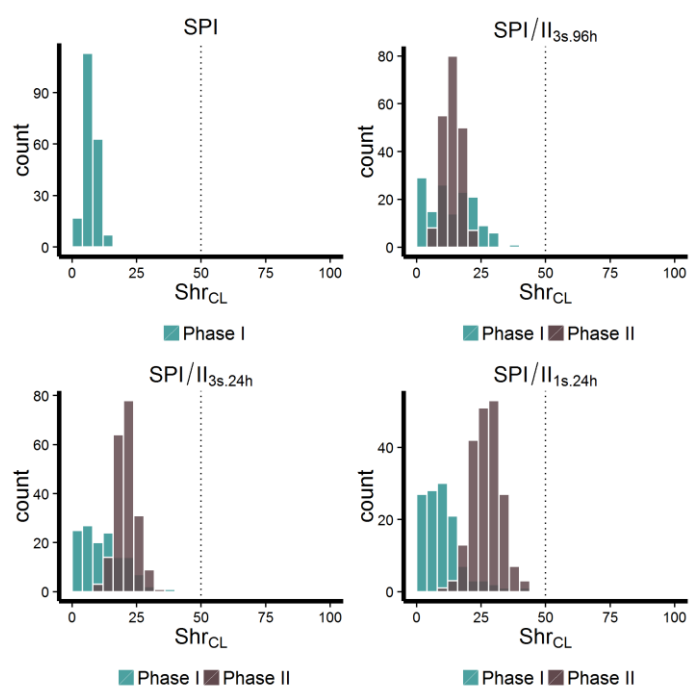
Total number of true positives (TP) with their 95% confidence interval under the alternative hypothesis.

On 200 simulated data sets, overall 1200 SNPs were set to impact clearance (maximum TP number).

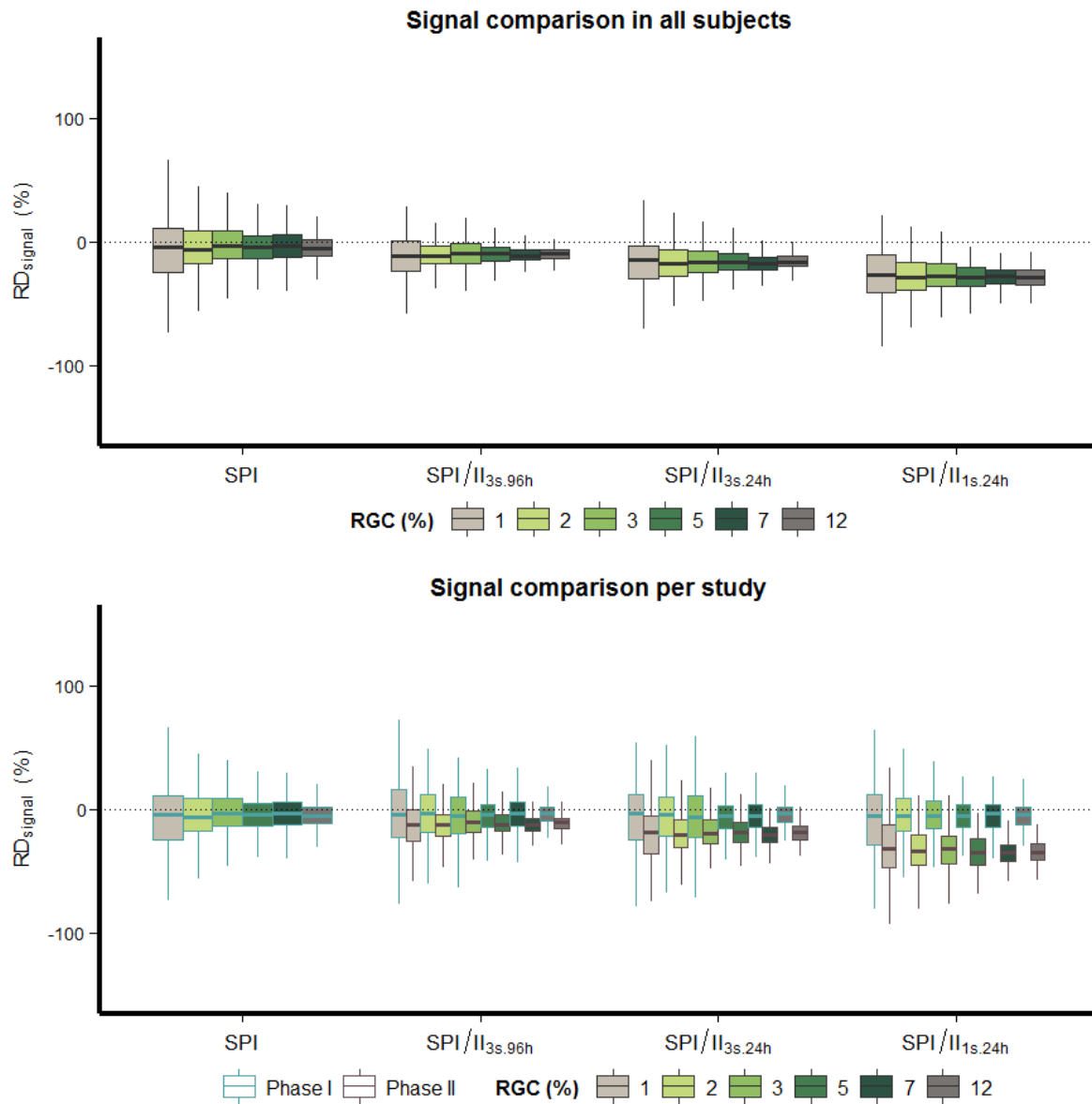
**Table S6.** False positives counts on CL, Q and V2 under  $H_1$  for Lasso and stepwise procedure methods in all scenarios simulated with  $IIV_{CL} = 60\%$ .

Method	SPI <sub>60%</sub>			SPI/II <sub>3s.96h_60%</sub>			SPI/II <sub>3s.24h_60%</sub>			SPI/II <sub>1s.24h_60%</sub>		
	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)
Lasso	34 [24;48]	43 [31;58]	15 [8;25]	171 [146;199]	28 [19;40]	29 [19;42]	161 [137;188]	117 [97;140]	39 [28;53]	150 [127;176]	81 [64;101]	24 [15;36]
Stepwise procedure	28 [19;40]	44 [32;59]	12 [6;21]	99 [80;121]	31 [21;44]	27 [18;39]	97 [79;118]	111 [91;134]	66 [51;84]	106 [87;128]	82 [65;102]	36 [25;50]

Total number of false positives (FP) with their 95% confidence interval under the alternative hypothesis.



**Figure S7.** Distribution of the shrinkages on clearance for subjects in the phase I dataset (blue) and for subjects in the phase II dataset (brown), for each scenario simulated under  $H_0$  with  $IIV_{CL} = 60\%$ .



**Figure S8.** Boxplot showing the loss of the signal for genetic effect in the overall population (*top*), as well as separately for the phase I data (blue borders) and for the phase II data (brown borders) (*bottom*). A boxplot is shown as a function of increasing  $R_{GC}$  (boxplots colour) separately for each scenario simulated under  $H_1$  with  $IIV_{CL} = 60\%$ .

## Additional results for the two other penalised regression methods

Ridge regression<sup>12</sup> imposes a penalty on the size of the  $\beta_k$  to reduce the prediction error without preventing the inclusion of variables in the model, by applying a Gaussian prior of identical variance on the eigenvalues issued from the principal component analysis (PCA) of the data. We used the approach proposed by Cule et al.<sup>10</sup> to semi-automatically set the penalty. The fit was followed by a Wald test on these coefficients and their standard error to perform the variable selection<sup>13</sup>, with a significance threshold equal to  $\alpha$ , the type I error per SNP.

HyperLasso<sup>11</sup> is similar to Lasso but uses a normal-exponential gamma (NEG) distribution as a prior on  $\beta_k$  and depends on two parameters: a shape parameter  $\lambda$  and a scale parameter  $\gamma$ . The sharp peak at zero and the flatter tail of the NEG distribution favour sparse solutions but the estimates of larger effects are shrunk less severely than the Lasso. The smaller the shape parameter the heavier the tails of the distribution and the more peaked at zero, which can result in fewer correlated SNPs being selected. The shape parameter  $\lambda$  was set to 1, which gives realistic effect size distributions<sup>14</sup>. The scale parameter  $\gamma$  was computed, as for the Lasso tuning parameter, depending on  $\alpha$ <sup>11</sup>.

**Table S7.** Empirical estimates of Family-Wise Error Rate under  $H_0$  for ridge regression and HyperLasso methods

Method		FWER (%)			
		SPI	SPI/II <sub>3s,96h</sub>	SPI/II <sub>3s,24h</sub>	SPI/II <sub>1s,24h</sub>
Ridge regression	Without correction <sup>a</sup>	16	15.5	18.5	14
HyperLasso	Without correction <sup>a</sup>	13.5	18	21.5	13.5
Ridge regression	After empirical correction <sup>b</sup>	21	20.5	18.5	19.5
HyperLasso	After empirical correction <sup>b</sup>	20.5	20	21.5	18

a. Set of empirical family wise error rates (FWER) obtain without correction.

b. Set of empirical FWER obtained after correction of thresholds or penalisation parameters. The 95% prediction interval around 20 for 200 simulated data sets is [14.5-25.5].

**Table S8.** Counts of true positives under  $H_1$  for ridge regression and HyperLasso methods in all scenarios

Method	SPI	TP (CL)		
		SPI/II <sub>3s,96h</sub>	SPI/II <sub>3s,24h</sub>	SPI/II <sub>1s,24h</sub>
Ridge regression	57 [43;74]	369 [332;409]	300 [267;336]	240 [211;272]
HyperLasso	41 [29;56]	309 [276;345]	250 [220;283]	225 [197;256]

Total number of true positives (TP) with their 95% confidence interval under the alternative hypothesis.

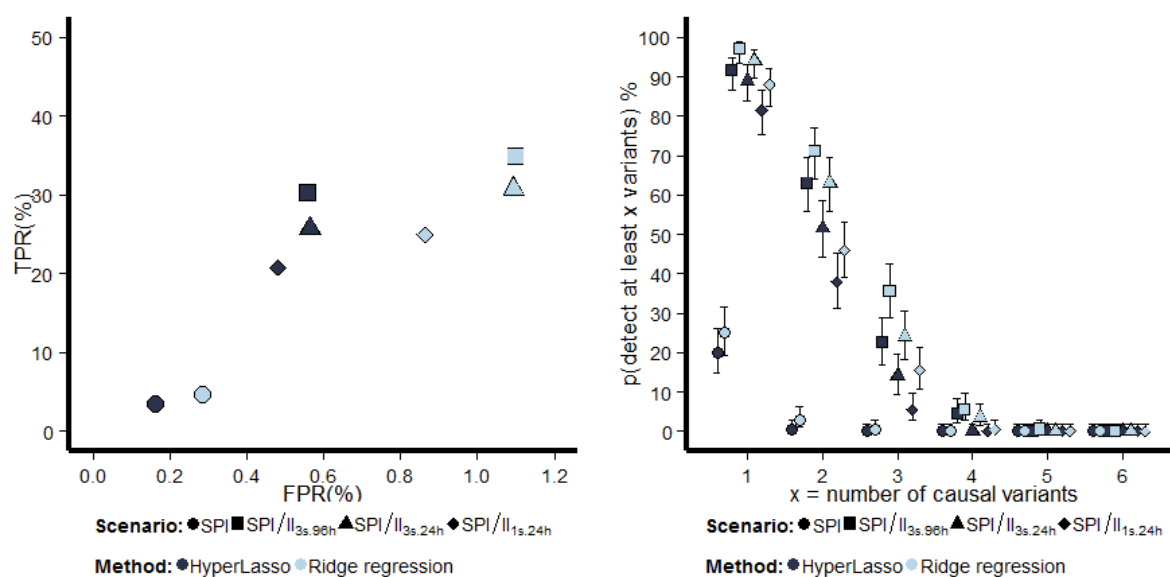
On 200 simulated data sets, overall 1200 SNPs were set to impact clearance (maximum TP number).



**Table S9.** Counts of false positives on CL, Q and V2 under  $H_1$  regression and HyperLasso methods in all scenarios

Method	SPI			SPI/II <sub>3s.96h</sub>			SPI/II <sub>3s.24h</sub>			SPI/II <sub>1s.24h</sub>		
	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)
Ridge regression	45	42	10	247	48	79	201	136	35	179	79	36
	[33;60]	[30;57]	[5;18]	[217;280]	[35;64]	[63;98]	[174;231]	[114;161]	[24;49]	[154;207]	[63;98]	[25;50]
HyperLasso	15	33	7	88	38	64	72	93	27	78	59	27
	[8;25]	[23;46]	[3;14]	[71;108]	[27;52]	[49;82]	[56;91]	[75;114]	[18;39]	[62;97]	[45;76]	[18;39]

Total number of false positives (FP) with their 95% confidence interval under the alternative hypothesis.



**Figure S9.** True Positive Rate (TPR) versus False Positive Rate (FPR) under  $H_1$  (*left*) and probability estimates (points) and 95% confidence interval (bars) to detect at least  $x$  causal variants explaining the interindividual variability of CL ( $x=1,\dots,6$ ) under  $H_1$  (*right*). Different symbols are used for each scenario, and colours denote the HyperLasso (grey) and the ridge regression (light blue).

## Additional references

1. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
2. Kuhn, E. & Lavielle, M. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat.* **8**, 115–131 (2004).
3. Sheiner, L. B., Rosenberg, B. & Melmon, K. L. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput. Biomed. Res. Int. J.* **5**, 411–459 (1972).
4. Hill, W. G. & Robertson, A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).
5. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
6. Bertrand, J., Comets, E., Laffont, C. M., Chenel, M. & Mentré, F. Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm. *J. Pharmacokinet. Pharmacodyn.* **36**, 317–339 (2009).
7. Savic, R. M. & Karlsson, M. O. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *AAPS J.* **11**, 558–569 (2009).
8. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).
9. R Development Core Team *A language and environment for statistical computing. R Foundation for Statistical Computing.* (2012).at <<http://www.R-project.org/>>
10. Cule, E. & De Iorio, M. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet. Epidemiol.* **37**, 704–714 (2013).
11. Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
12. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55 (1970).
13. Cule, E., Vineis, P. & De Iorio, M. Significance testing in ridge regression for genetic data. *BMC Bioinformatics* **12**, 372 (2011).
14. Vignal, C. M., Bansal, A. T. & Balding, D. J. Using penalised logistic regression to fine map HLA variants for rheumatoid arthritis. *Ann. Hum. Genet.* **75**, 655–664 (2011).

# Chapitre 6

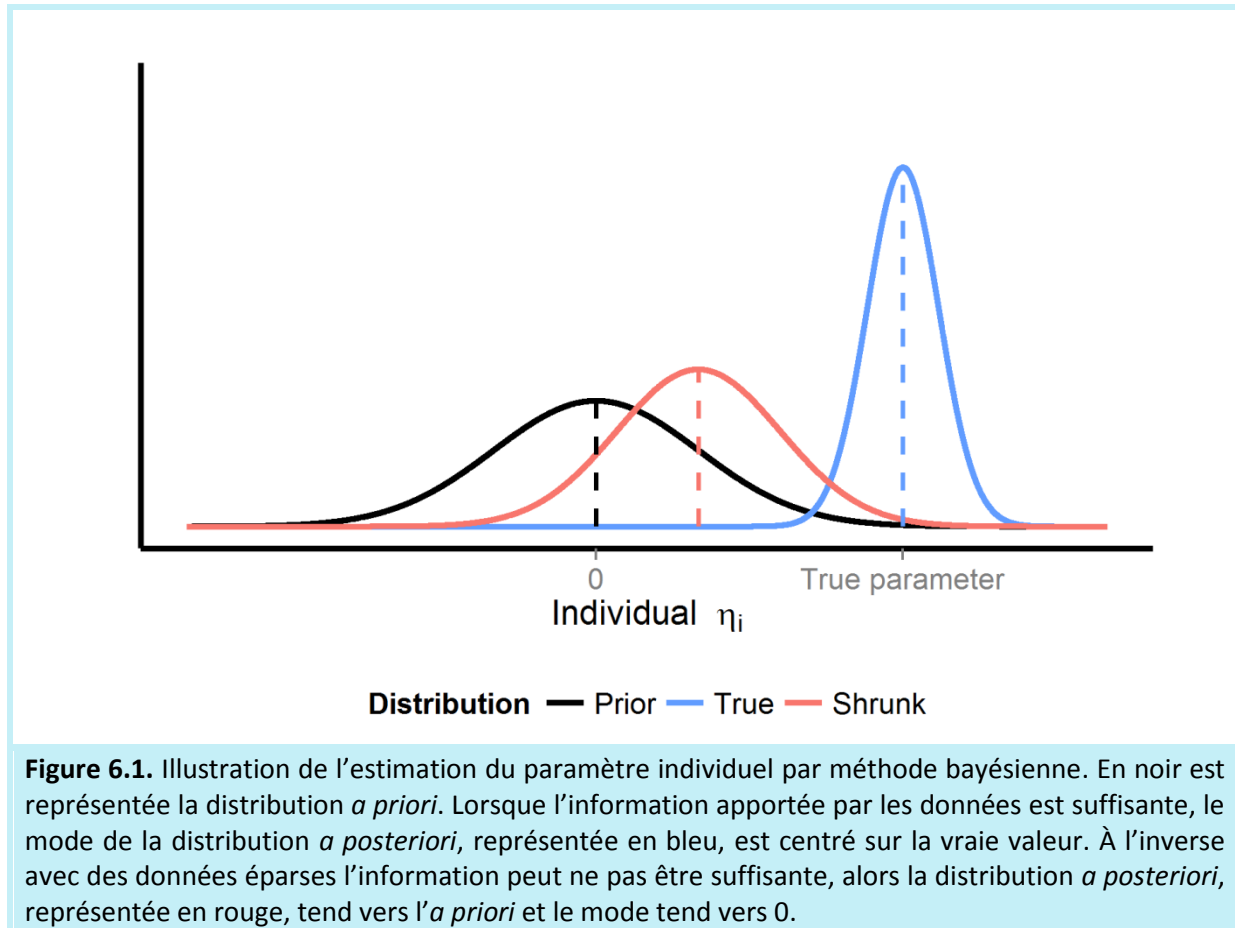
## PRISE EN COMPTE DE L'INTÉGRALITÉ DE LA DISTRIBUTION CONDITIONNELLE DANS LES TESTS D'ASSOCIATION

### 6.1. Introduction

Dans le premier travail présenté au chapitre 4, nous constatons la faible puissance de détection des effets génétiques dans une étude de phase I typique, en raison du faible nombre de sujets. Nous explorons ainsi deux pistes pour améliorer cette puissance. Tout d'abord nous proposons de combiner ces données de phase I à des données issues d'une étude de phase II, augmentant ainsi le nombre de sujets. Le chapitre 5 montre l'intérêt de cette approche se basant sur des protocoles mixtes dans l'augmentation de la puissance de détection des polymorphismes génétiques. Dans ce chapitre, nous proposons d'évaluer si la prise en compte de toute la distribution des paramètres individuels permet d'améliorer la puissance des méthodes de détection.

Nous avons montré le lien entre la puissance de détection des effets gènes et la quantité d'information permettant d'estimer le paramètre d'intérêt, que l'on peut mesurer par le  $\eta$ -*shrinkage* sur l'estimation des paramètres individuels (EBE). L'estimation des paramètres individuels se fait par une méthode bayésienne utilisant la distribution *a priori* du paramètre estimée lors de l'analyse de population et les données d'un sujet afin d'estimer une distribution *a posteriori* (voir paragraphe 1.1.5.3). Le mode de cette distribution est généralement utilisé dans les différents graphiques diagnostics, et également lors de la pré-sélection des covariables. Ce mode tend vers la valeur de population (*shrinkage*) lorsque l'information du sujet n'est pas suffisante (Savic and Karlsson, 2009), en raison de données éparées (**Figure 6.1**). Une alternative pourrait être de prendre en compte, non pas uniquement cette valeur ponctuelle, mais l'ensemble de la distribution *a posteriori*. Dans les dernières versions du logiciel Monolix ([www.lixoft.eu](http://www.lixoft.eu)) des tirages dans la distribution conditionnelle des paramètres individuels sont utilisés dans certains graphiques d'évaluation du modèle développé et notamment dans les graphiques représentant i) les paramètres

individuels en fonction des covariables et ii) la distribution des paramètres individuels et le  $\eta$ -shrinkage associé. Ces tirages permettent de décrire la distribution *a posteriori* entière et seraient donc moins influencés que le mode par une information éparse.



Nous nous proposons d'utiliser cette nouvelle approche dans les tests d'association afin d'augmenter la probabilité de détecter des variants génétiques, notamment dans le cas d'études avec un protocole épars, où l'effet du  $\eta$ -shrinkage diminue la puissance des tests. Nous évaluons cette approche dans une étude de simulation.

## 6.2. Matériels et méthodes

### 6.2.1. Étude de simulation

Les simulations se basent à nouveau sur le cas réel présenté au chapitre 2. Le modèle PK développé pour cet exemple est utilisé pour simuler des concentrations sous  $H_0$  et  $H_1$ , en prenant en compte les génotypes simulés par ailleurs pour 176 SNPs. À la différence des travaux présentés dans les chapitres 4 et 5, seul un polymorphisme affecte la

pharmacocinétique de la molécule sous  $H_1$ , afin de simplifier l'analyse en n'ayant à considérer que des régressions univariées. Ce polymorphisme causal est tiré aléatoirement pour chaque jeu de données à partir des 176 SNPs simulés et affecte la clairance avec un effet additif, comme représenté ci-dessous :

$$\log(CL_i) = \log(\mu_{CL}) + \beta_s \times SNP_{is} + \eta_{CL_i} \quad (6.1)$$

où  $CL_i$  est la clairance individuelle,  $\mu_{CL}$  la valeur typique de CL dans la population,  $\beta_s$  l'effet associé au polymorphisme  $SNP_{is}$  et  $\eta_{i_{CL}}$  l'effet aléatoire pour le sujet  $i$ . L'effet  $\beta_s$  du SNP causal sur CL dépend de deux paramètres : la part de la variabilité de CL expliquée par le polymorphisme ( $R_{GC_s}$  exprimé en pourcentage) et la fréquence de l'allèle muté pour le variant génétique, ou *Minor allele fraction* ( $MAF_s$ ) (Bertrand and Balding, 2013). Le coefficient  $\beta_s$  est calculé selon la formule :

$$\beta_s = \sqrt{\frac{R_{GC_s} \times \omega_{CL}^2}{2MAF_s(1 - MAF_s) - R_{GC_s} \times 2MAF_s(1 - MAF_s)}} \quad (6.2)$$

où  $\omega_{CL}^2$  représente la variance des effets aléatoires de CL due à des sources non génétiques. Différents scénarios sont simulés, avec trois tailles d'effet pour le variant causal, quantifiés par le paramètre  $R_{GC_s}$  (5, 12 et 20%). Pour chaque scénario, 78 sujets sont simulés selon deux protocoles de prélèvement. Soit l'ensemble des sujets est simulé avec un protocole riche comprenant 16 points de prélèvements, soit avec un protocole épars où chaque sujet présente une observation. Pour permettre, malgré le protocole très épars, une estimation précise des paramètres de population, les sujets sont divisés en trois groupes de 26 sujets, chaque groupe étant associé à un temps de prélèvement différent. En tout, 6 scénarios sont donc évalués (**Table 6.1**) et au total 100 jeux de données sont simulés pour chaque scénario.

**Table 6.1.** Scénarios simulés.

Protocole	$R_{GC} = 5\%$	$R_{GC} = 12\%$	$R_{GC} = 20\%$
Riche <sup>a</sup>	$S_{5\%.rich}$	$S_{12\%.rich}$	$S_{20\%.rich}$
Épars <sup>b</sup>	$S_{5\%.sparse}$	$S_{12\%.sparse}$	$S_{20\%.sparse}$

a. Seize observations par sujet (0,5, 1, 1,5, 2, 3, 4, 6, 8, 12, 16, 24, 48, 72, 96, 120 et 192h).

b. Une observation par sujet, trois groupes (2, 22 ou 192h).

À partir des profils PK simulés pour chaque scénario et avec le même modèle utilisé pour les simulations, les paramètres individuels (EBE) sont ré-estimés à l'aide de l'algorithme SAEM

(Kuhn and Lavielle, 2004) et du logiciel Monolix version 4.3.1 ([www.lixoft.eu](http://www.lixoft.eu)), accompagné du logiciel MATLAB ([www.mathworks.com](http://www.mathworks.com)). Pour un sujet  $i$ , l'EBE de la clairance correspond à l'effet aléatoire  $\hat{\eta}_{CL_i}$ , i.e. le *Maximum A posteriori* (MAP) de la distribution conditionnelle. En plus de l'estimation du mode,  $k$  tirages sont effectués aléatoirement dans la distribution conditionnelle du paramètre individuel  $\eta_{CL_i}$  (**Figure 6.1**).

Nous comparons ensuite l'utilisation de ces deux variables comme phénotype dans les analyses d'association avec le variant causal, selon deux modèles. Comme précédemment, un modèle linéaire simple est d'abord utilisé pour tester l'association entre le SNP causal et les estimations individuelles des effets aléatoires de la clairance  $\hat{\eta}_{CL_i}$  (**équation 6.3**), où  $\beta_0$  représente l'intercept,  $\beta_s$  le coefficient d'effet associé au variant causal  $SNP_{is}$ , et  $\varepsilon_i$  une erreur Gaussienne.

Un modèle linéaire à effets mixtes est par contre utilisé pour tester l'association entre le SNP causal et les tirages à partir de la distribution conditionnelle de CL ( $\eta_{CL_{i,k}}^*$ ), où  $\delta_k$  représente l'effet aléatoire quantifiant la différence entre les différents tirages pour un même individu afin de tenir compte de la corrélation entre ces tirages, et  $\varepsilon_{ik}$  à nouveau une erreur Gaussienne (**équation 6.4**).

$$\hat{\eta}_{CL_i} = \beta_0 + \beta_s \times SNP_{is} + \varepsilon_i \quad (6.3)$$

$$\eta_{CL_{i,k}}^* = \beta_0 + \beta_s \times SNP_{is} + \delta_k + \varepsilon_{ik} \quad (6.4)$$

Afin de tester l'effet du nombre de tirages  $k$ , cette procédure a été réalisée avec 3, 10, 100, 300 ou 600 tirages.

### 6.2.2. Obtention des tirages dans la distribution conditionnelle

Il s'avère que le nombre de tirages ne peut être fixé par l'utilisateur dans Monolix et qu'il est, de plus, très limité. Afin d'augmenter le nombre de tirages dans la distribution conditionnelle pour chaque jeu de données simulé, un algorithme a été développé pour permettre 100 tirages aléatoires de  $k$  valeurs par jeu de données :

1. Estimation des paramètres de population
2. Pour 100 itérations :
  - i. Une nouvelle graine est fixée
  - ii. Nouvelle estimation des paramètres de population  
Valeurs initiales fixées aux estimations de l'étape 1  
Uniquement une itération de SAEM pour l'estimation :  
→ estimations = estimations de l'étape 1
  - iii. Estimation des paramètres individuels ( $\hat{\eta}_{CL_i}$ )

- iv.  $k$  tirages dans la distribution conditionnelle de CL ( $\eta_{CL_i,k}^*$ )
- 3. Sauvegarde des estimations des paramètres de population, des estimations individuelles et des tirages concaténés à partir des 100 itérations

En changeant la graine à chaque itération et en ne permettant qu'une itération à l'algorithme SAEM pour l'estimation des paramètres de populations, les  $\hat{\eta}_{CL_i}$  ne sont pas modifiés d'une itération à l'autre de notre algorithme, tandis qu'on obtient des valeurs différentes pour les tirages  $\eta_{CL_i,k}^*$ . Ainsi le nombre de tirages est augmenté jusqu'à 600 pour chaque jeu de données.

### 6.2.3. Évaluation

Sous  $H_0$ ,  $\eta$ -shrinkage est quantifié selon la formule proposée par Bertrand et al (Bertrand et al., 2009), comparant la variance des effets aléatoires ( $\hat{\eta}_{CL_i}$ ) à la variance de la distribution  $a$  priori calculée à l'étape de population (paragraphe 1.1.5.3, **équation 1.14**). Pour vérifier que les tirages décrivent bien la variabilité de la distribution, nous pouvons calculer un équivalent (que nous appelons *sample-shrinkage*) en utilisant la variance de ces  $k$  tirages, comparée à la variance de la distribution  $a$  priori, selon la formule :

$$sample-Sh_{CL} = 1 - \frac{var(\eta_{CL_i,k}^*)}{\hat{\omega}^2} \quad (6.5)$$

Toujours sous  $H_0$ , l'erreur de type I est calculée comme le pourcentage de jeux de données où l'association entre le SNP causal et le phénotype est significative. Du fait que seule cette association est testée, aucune correction pour test multiple n'est utilisée pour ce travail. L'erreur de type I doit être contrôlée autour de 5%, avec un intervalle de prédiction à 95% pour 100 jeux de données égal à [0,73 ; 9,27] %.

Sous  $H_1$ , la probabilité de détection est quantifiée comme étant le pourcentage de jeux de données où le SNPs causal est significativement associé à  $\hat{\eta}_{CL_i}$  ou  $\eta_{CL_i,k}^*$ , selon le modèle utilisé.

## 6.3. Résultats

### 6.3.1. Erreur de type I

L'erreur de type I estimée sur 100 jeux de données dans chaque scénario est correctement contrôlée autour de 5% (**Table 6.2**), aussi bien lorsque le test de covariable est effectué sous  $H_0$  sur les estimations individuelles  $\hat{\eta}_{CL_i}$ , que sur les tirages dans la distribution



conditionnelle  $\eta_{CL_i,k}^*$ , et ce indépendamment du nombre de tirages. En effet les valeurs de l'erreur de type I s'échelonnent entre 3 et 8% et sont donc comprises dans l'intervalle de prédiction à 95% autour de la valeur cible de 5%. Aussi pour le calcul de la puissance sous  $H_1$ , aucune correction n'est effectuée.

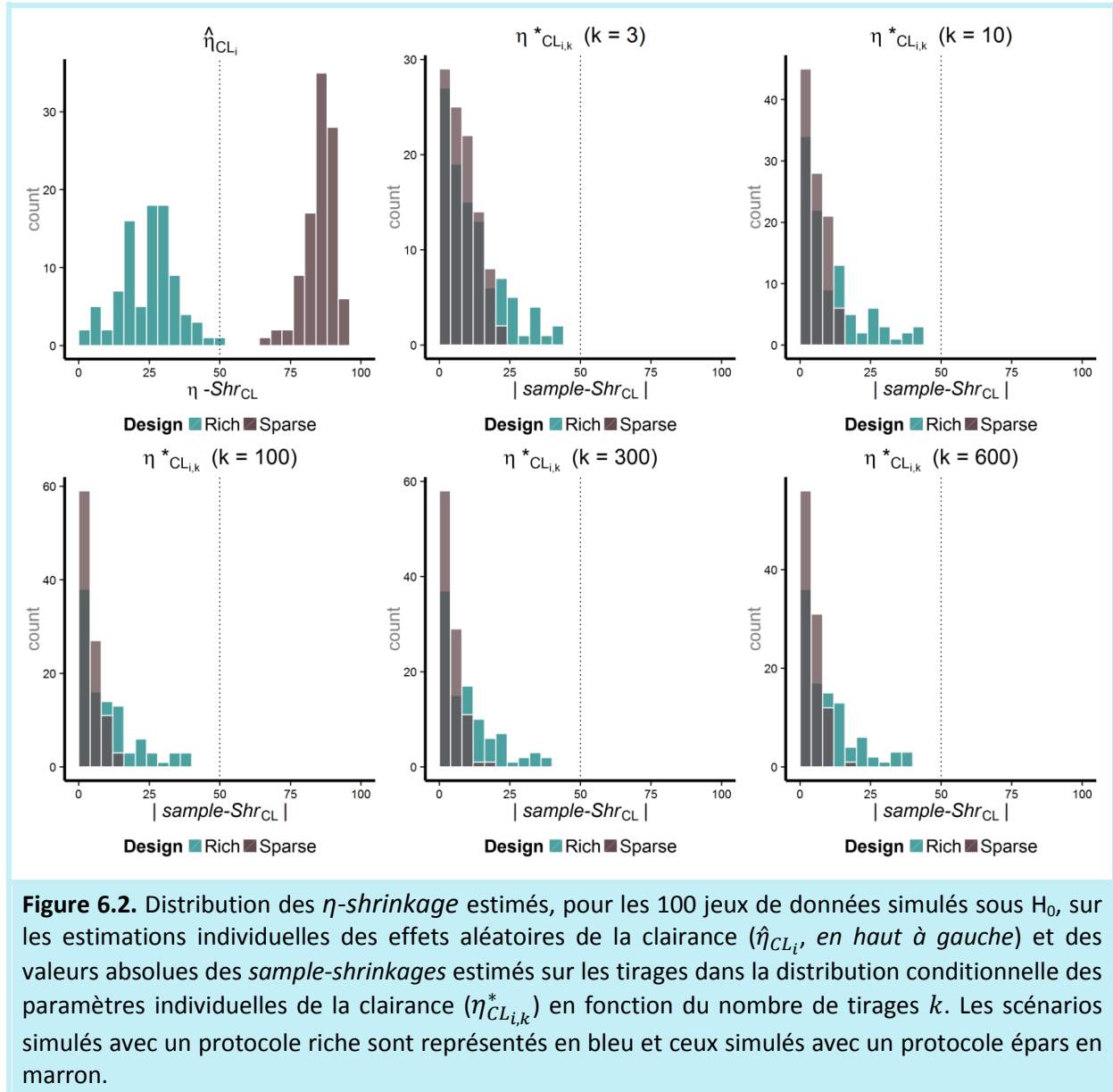
**Table 6.2.** Estimation empirique de l'erreur de type I sous  $H_0$  pour les différents scénarios.

Phenotype	Type I error (%)					
	S5%.rich	S12%.rich	S20%.rich	S5%.sparse	S12%.sparse	S20%.sparse
$\hat{\eta}_{CL_i}$	8,0	8,0	8,0	3,0	7,0	7,0
$\eta_{CL_i,k}^* (k = 3)$	5,0	6,0	6,0	5,0	5,0	5,0
$\eta_{CL_i,k}^* (k = 10)$	6,0	7,0	7,0	4,0	5,0	5,0
$\eta_{CL_i,k}^* (k = 100)$	8,0	8,0	8,0	4,0	6,0	6,0
$\eta_{CL_i,k}^* (k = 300)$	8,0	8,0	8,0	4,0	7,0	7,0
$\eta_{CL_i,k}^* (k = 600)$	8,0	8,0	8,0	4,0	7,0	7,0

L'intervalle de prédiction à 95% autour de 5 pour 100 jeux de données simulés est [0,73 ; 9,27]%.

### 6.3.2. Shrinkage

Comme attendu avec les protocoles choisis pour simuler les profils individuels, le  $\eta$ -shrinkage calculé sur la variance des estimations individuelles  $\hat{\eta}_{CL_i}$  est limité dans les simulations avec un protocole riche, la majorité des jeux de données présentant un  $\eta$ -shrinkage inférieur à 50% (**Figure 6.2**). À l'inverse, pour les simulations avec un protocole épars l'étendue du  $\eta$ -shrinkage calculé sur les estimations individuelles est grande et ce dernier est supérieur à 50% pour tous les jeux de données.



**Figure 6.2.** Distribution des  $\eta$ -shrinkage estimés, pour les 100 jeux de données simulés sous  $H_0$ , sur les estimations individuelles des effets aléatoires de la clairance ( $\hat{\eta}_{CL_i}$ , en haut à gauche) et des valeurs absolues des  $\text{sample-shrinkage}$  estimés sur les tirages dans la distribution conditionnelle des paramètres individuelles de la clairance ( $\eta^*_{CL_{i,k}}$ ) en fonction du nombre de tirages  $k$ . Les scénarios simulés avec un protocole riche sont représentés en bleu et ceux simulés avec un protocole épars en marron.

Le  $\text{sample-shrinkage}$  calculé sur la variance des tirages dans la distribution conditionnelle de CL se répartit de manière homogène autour de 0. Représenté en valeur absolue (**Figure 6.2**), le  $\text{sample-shrinkage}$ , comparé au  $\eta$ -shrinkage, diminue pour les jeux de données simulés avec un protocole riche et diminue fortement pour ceux simulés avec un protocole épars. Le  $\text{sample-shrinkage}$  sur les échantillons de la distribution conditionnelle est alors inférieur à 50% pour tous les jeux de données et ce quel que soit le nombre de tirages utilisé. Ce résultat laisse à penser que l'on décrirait mieux la variabilité interindividuelle en utilisant la distribution conditionnelle, surtout dans le cas de protocoles épars associés à un fort  $\eta$ -shrinkage sur le MAP.

### 6.3.3. Probabilité de detection

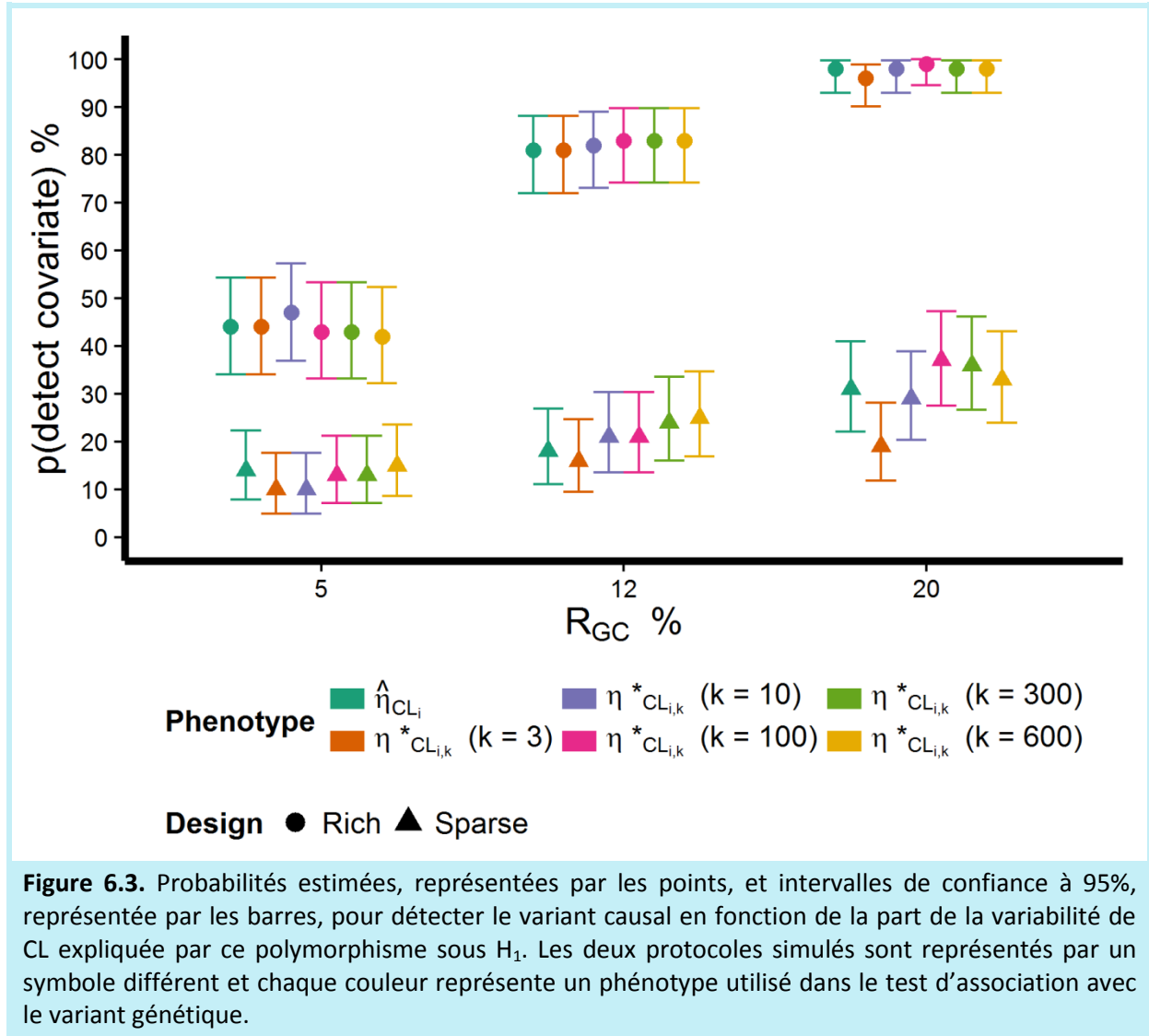
Comme attendu, le nombre de jeux de données dans lesquels le variant causal est détecté augmente avec l'effet (**Table 6.3**). De même un protocole riche permet de détecter l'effet de la génétique dans un plus grand nombre de jeux de données qu'un protocole épars, peu informatif.

**Table 6.3.** Nombre de jeux de données N simulés sous  $H_1$  où le variant causal a été détecté.

Phenotype	N					
	$S_{5\%.rich}$	$S_{12\%.rich}$	$S_{20\%.rich}$	$S_{5\%.sparse}$	$S_{12\%.sparse}$	$S_{20\%.sparse}$
$\hat{\eta}_{CL_i}$	44 [34;54]	81 [72;88]	98 [93;100]	14 [8;22]	18 [11;27]	31 [22;41]
<i>sample</i> <sub><math>ik_{\eta_{CL}}</math></sub> (k = 3)	44 [34;54]	81 [72;88]	96 [90;99]	10 [5;18]	16 [9;25]	19 [12;28]
<i>sample</i> <sub><math>ik_{\eta_{CL}}</math></sub> (k = 10)	47 [37;57]	82 [73;89]	98 [93;100]	10 [5;18]	21 [13;30]	29 [20;39]
<i>sample</i> <sub><math>ik_{\eta_{CL}}</math></sub> (k = 100)	43 [33;53]	83 [74;90]	99 [95;100]	13 [7;21]	21 [13;30]	37 [28;47]
<i>sample</i> <sub><math>ik_{\eta_{CL}}</math></sub> (k = 300)	43 [33;53]	83 [74;90]	98 [93;100]	13 [7;21]	24 [16;34]	36 [27;46]
<i>sample</i> <sub><math>ik_{\eta_{CL}}</math></sub> (k = 600)	42 [32;52]	83 [74;90]	98 [93;100]	15 [9;24]	25 [17;35]	33 [24;43]

Nombre de jeux de données et son intervalle de confiance à 95%.  
Le nombre maximal pouvant être atteint par N est 100.

Ces différences se retrouvent dans la probabilité de détecter le SNP causal dans les 100 jeux de données (**Figure 6.3**). Dans le cas des simulations avec un protocole riche, l'utilisation des tirages dans la distribution conditionnelle de CL ne permet pas d'améliorer la probabilité de détecter le variant génétique, par rapport à l'utilisation des estimations des effets aléatoires. Cette probabilité ne varie pas de plus de 3 points, et ce quel que soit le nombre de tirages utilisés. Pour les protocoles épars, des différences plus importantes mais toujours très limitées sont observées. L'augmentation de la probabilité de détection grâce à l'utilisation des tirages dans la distribution conditionnelle vaut au maximum 1, 7 et 6 points respectivement pour les scénarios simulés avec un  $R_{GC_S}$  égal à 5, 12 et 20%. Un nombre important de tirages (100 à 600) est de plus nécessaire pour observer ces améliorations.



## 6.4. Discussion

Dans ce travail, nous cherchons à évaluer par une étude de simulation l'utilisation d'une nouvelle approche basée sur des tirages aléatoires dans la distribution conditionnelle des paramètres individuels. Les estimations individuelles des paramètres du modèle sont généralement obtenues à partir de l'estimateur du MAP, *i.e.* le mode de la distribution conditionnelle. Ces estimations individuelles sont ensuite utilisées dans les évaluations graphiques, représentant notamment les relations EBE - covariables, et dans les tests de covariables. Le MAP est sensible au  $\eta$ -shrinkage entraînant une perte d'information sur la variabilité interindividuelle des paramètres et donc des biais dans l'évaluation graphique et les tests de covariables. L'échantillonnage dans la distribution conditionnelle des paramètres individuels permet d'utiliser toute la distribution, et non plus uniquement son mode. Cela

permettrait de mieux quantifier la variabilité interindividuelle, et donc de mieux représenter ou tester les relations paramètres - covariables.

Dans le cas du protocole riche, l'utilisation des tirages dans la distribution conditionnelle de CL n'apporte pas plus de puissance pour détecter le variant causal, comparé à l'utilisation du mode des estimations individuelles des effets aléatoires de CL. Ce résultat est attendu car le  $\eta$ -shrinkage sur les estimations individuelles de CL est limité du fait du nombre important d'observations par sujet. Cette approche permettant potentiellement de corriger les effets du shrinkage ne peut logiquement affecter que de façon très limitée la puissance de détection pour des protocoles riches. À l'inverse on observe une amélioration de cette puissance avec l'utilisation des tirages dans la distribution conditionnelle quand le nombre d'observations par sujet est réduit, contrant en partie le fort  $\eta$ -shrinkage sur les estimations individuelles de CL. Mais cette amélioration reste limitée, la probabilité de détecter le variant causal augmentant au mieux de 7 points, et n'est observée que pour un grand nombre de tirages (au moins égal à 100). Cette légère augmentation de la puissance de détection ne peut être reliée à la valeur du *sample-shrinkage*, calculé sur les tirages de la distribution conditionnelle, qui est très inférieure au  $\eta$ -shrinkage, calculé sur les estimations individuelles. La faible magnitude du *sample-shrinkage* suggère que cette approche permet effectivement de mieux quantifier la variabilité interindividuelle avec des protocoles épars mais n'améliore pas la perte du signal génétique, la distribution conditionnelle étant toujours centrée sur des EBE subissant un shrinkage. L'utilisation de cette méthode semble donc surtout utile pour une description graphique de la variabilité interindividuelle, comme cela est fait actuellement dans le logiciel Monolix. À l'inverse, l'apport de cette méthode pour améliorer la détection des effets génétiques dans les tests de covariables est limité.

Des difficultés de mise en place de cette méthode ont de plus dû être contournées. En effet le nombre de tirages effectués dans la distribution conditionnelle ne peut être fixé *via* le logiciel Monolix. Afin d'obtenir un grand nombre de tirages, un algorithme est utilisé pour répéter l'étape d'estimation et de tirage sur un même jeu de données. Ce travail montre que la différence de puissance pour détecter le variant génétique entre les tests utilisés varie en fonction du nombre de tirages. Un nombre limité de simulations (100 jeux de données pour chaque scénario) a été effectué du fait de très longs temps de calcul. Cela est lié à la nécessité de ré-estimer plusieurs fois les paramètres (100 fois) sur un même jeu de données

afin d'augmenter le nombre de tirages. L'utilisation d'un nombre plus important de simulations permettrait de mieux estimer l'impact de l'utilisation des tirages dans la distribution conditionnelle des paramètres individuels sur la puissance de détection du variant causal. Mais les résultats ne devraient pas être modifiés de façon significative et les conclusions de ce travail devraient rester valables.

À notre connaissance, il n'est pas possible à l'heure actuelle d'obtenir ces tirages dans le logiciel NONMEM ([www.icon.com](http://www.icon.com)) alors que le package « saemix » (Comets et al., 2011) développé pour le logiciel R (R Development Core Team, 2012) intègre cette fonction. Par ailleurs les tests d'association présentés dans les chapitres précédents et basés sur des régressions pénalisées (la régression *ridge* (Cule and De Iorio, 2013), Lasso (Tibshirani, 1994), HyperLasso (Hoggart et al., 2008)) ne permettent pas d'intégrer un effet aléatoire afin de prendre en compte la corrélation entre les différents tirages dans un modèle linéaire à effets mixtes. Seuls des modèles linéaires simples sont à ce jour implémentés dans les logiciels ou packages permettant d'utiliser ces méthodes et l'utilisation de modèles mixtes complexifierait grandement le calcul d'une vraisemblance pénalisée. À l'inverse il n'y a pas d'obstacle à la mise en œuvre de modèle linéaire à effets mixtes dans une approche *stepwise procedure*, utilisée dans les travaux précédents.

Pour conclure, cette nouvelle approche pour les tests de covariables utilisant des tirages dans la distribution conditionnelle des paramètres individuels améliore de façon très modérée la probabilité de détecter un variant génétique dans le contexte de protocoles épars, où le *shrinkage* est élevé. L'apport de cette nouvelle approche est d'autant plus limité que les résultats présentés dans le chapitre 5 montraient une plus grande amélioration de la puissance des tests de détection en optimisant le protocole épars des études de phase II. D'après ces deux travaux, Il serait plus intéressant d'optimiser les études en amont afin de limiter le *shrinkage*, plutôt que de le corriger *a posteriori via* cette nouvelle approche.

Ce travail a fait l'objet d'une communication orale au 24<sup>ème</sup> congrès du PAGE (*Population Approach Group in Europe*), en juin 2015. Cette communication reprenait également les travaux présentés aux chapitres 4 et 5.

# Chapitre 7

## DISCUSSION GÉNÉRALE

### 7.1. Discussion des résultats

L'objectif premier de cette thèse était d'évaluer la méthodologie des analyses PGt, *i.e.* aussi bien la conception des études que les méthodes elles-mêmes, afin d'avoir suffisamment de puissance pour mettre en évidence l'effet de polymorphismes génétiques sur la pharmacocinétique des médicaments. Dans un deuxième temps, nous voulions fournir des recommandations afin d'améliorer cette puissance de détection. Pour cela plusieurs séries de simulations ont été effectuées, qui ont comme originalité de se baser sur un exemple de données réelles pour une molécule en cours de développement. Le protocole d'études de phase I, le modèle PK développé sur les données de ces études et la conception d'une puce à ADN ciblée sur l'étude des polymorphismes impliqués dans la pharmacocinétique des médicaments nous ont permis de simuler des conditions réalistes dans nos différents travaux. Différents résultats ressortent de ces travaux de simulations.

Une première comparaison concerne les méthodes d'estimation des paramètres PK d'une molécule, utilisés comme phénotype dans les analyses PGt. L'AUC et la clairance estimées par NCA et par modélisation étant corrélées, nous pouvions nous attendre à des probabilités similaires de détecter des polymorphismes génétiques sur ces deux phénotypes. Nous avons constaté que c'est en effet le cas lorsque le modèle PK est simplifié pour simuler une pharmacocinétique linéaire, sans variabilité interindividuelle sur l'absorption (en particulier sur la biodisponibilité). Cependant, lorsque la pharmacocinétique est non linéaire, la clairance estimée par modélisation permet une meilleure puissance de détection des variants causaux comparée à l'AUC estimée par NCA, et ce même en ajustant le phénotype NCA pour prendre en compte cette non linéarité. La NCA a l'avantage de ne pas reposer sur des hypothèses, qui peuvent orienter une analyse et donc les résultats. Mais cette approche demandant de nombreux prélèvements par sujet, elle n'est pas adaptée pour les phases plus avancées du développement clinique (phase II, phase III). La modélisation, en décomposant

les différentes phases du processus ADME, permet de mieux prendre en compte un facteur de variabilité comme l'effet de la dose sur l'absorption, comparé à la NCA qui estime des paramètres résumant l'ensemble des phases. La modélisation et l'approche de population ont également l'intérêt de pouvoir être utilisées avec des données éparées et des protocoles non homogènes, ce qui en fait des outils utilisables tout au long du développement d'un médicament.

Nous avons également comparé différentes méthodes testant les associations entre phénotypes PK et polymorphismes génétiques. Une approche itérative (*stepwise procedure*), où plusieurs étapes sont nécessaires pour obtenir un modèle final, était comparée à des méthodes basées sur des régressions pénalisées estimant et sélectionnant les variants causaux en une étape : la régression *ridge* associée à un test de Wald, le Lasso et l'HyperLasso. Un premier résultat concerne l'erreur de type I globale, inférieure au seuil ciblé. Afin de prendre en compte l'inflation de l'erreur de type I globale due au grand nombre de tests effectués, une approche basée sur le FWER a été utilisée dans les différents tests d'association. Une correction de Šidák était appliquée sur l'erreur de type I par test afin de contrôler le FWER, *i.e.* la probabilité de détecter au moins un faux positif. Les méthodes de correction de l'inflation de l'erreur de type I basées sur une correction de l'erreur de type I par test, comme la correction de Bonferroni ou de Šidák, ont tendance à sur-corriger les seuils de significativité quand les variables sont corrélées et que les différents tests sont supposés indépendants (Nyholt, 2004). Les polymorphismes génétiques, en raison du déséquilibre de liaison, sont corrélés et il peut être intéressant dans ce cas de prendre en compte la structure de la corrélation entre les tests (Bender and Lange, 2001). Un scénario où les polymorphismes étaient simulés sans corrélation confirme ce résultat, l'erreur de type I globale étant cette fois correctement contrôlée. Il était donc nécessaire dans nos travaux de corriger empiriquement une seconde fois les erreurs de type I par test pour obtenir le FWER ciblé. La correction de Bonferroni est, d'après la littérature, la méthode la plus utilisée dans les analyses PGt. En plus de la baisse de puissance de détection du polymorphisme qui accompagne une sur-corrrection de l'erreur de type I, la méthode de Bonferroni est connue pour être la plus conservatrice, aboutissant à des seuils de significativité par test extrêmement bas lorsque leur nombre est élevé. Ces seuils fortement corrigés dans les études d'association GWAS ou, dans une moindre mesure, basées sur une approche gènes candidats, conduisent à ne mettre en évidence que des variants fréquents ayant un fort effet



sur le phénotype. Ainsi, une partie des facteurs génétiques expliquant la variabilité du phénotype, de par leur effet plus modéré ou leur faible fréquence, sont indétectables avec des tailles de population réalistes. Il serait donc intéressant d'évaluer l'effet d'autres méthodes de correction sur la puissance de détection des analyses. Des méthodes alternatives sont à envisager, comme des méthodes basées sur les permutations (Bertrand, 2009), des méthodes de ré-échantillonnage qui ont l'inconvénient de nécessiter de longs temps de calculs.

En contrôlant la probabilité d'avoir au moins un faux positif parmi les relations sélectionnées, l'approche basée sur le FWER revient à sélectionner un nombre limité de variants lorsqu'un grand nombre d'associations sont testées, diminuant ainsi le nombre potentiel de vrais positifs. Cette approche est donc à l'origine d'une baisse de puissance des analyses. Une alternative à cette approche est d'autoriser un nombre plus grand de faux positifs, tout en s'assurant que ce nombre ne soit pas trop important. L'approche *False Discovery Rate* (FDR) permet de contrôler la proportion de faux positifs parmi l'ensemble des variants sélectionnés (Benjamini and Hochberg, 1995). Avec cette approche, on accepte que soient présentes dans le modèle final des variables dont l'effet est négligeable. Cela n'a pas d'effet sur l'adéquation du modèle aux données et, en matière de construction de modèles, ne va qu'à l'encontre du principe de parcimonie. Par contre, en acceptant plus de faux positifs, un nombre plus important d'associations sont mises en évidence, parmi lesquelles on trouve des associations réellement significatives. L'approche FDR permet donc une analyse plus sensible, mais moins spécifique. Elle pourrait donc être envisagée pour des analyses exploratoires. La contrepartie de cette approche est un nombre plus important de relations à confirmer dans des analyses ultérieures. Pour des études étudiant un nombre limité d'associations, le nombre de faux positifs resterait limité (en imaginant par exemple contrôler le taux de faux positifs à 5%). Pour des études à plus grandes échelles, le nombre de faux positifs serait plus important, impliquant une augmentation des coûts des études confirmatoires et des temps de développement. Enfin, l'objectif des analyses PGt est de mettre en évidence des associations pouvant conduire à des recommandations sur l'utilisation des médicaments en pratique clinique. Il n'y a donc pas d'intérêt à mettre en évidence des polymorphismes dont l'effet n'est pas cliniquement pertinent.

Le comportement des différents tests d'association évalués, en matière d'erreur de type I et de puissance de détection des polymorphismes génétiques, est similaire entre les méthodes

selon les différentes conditions de simulations (nombre de sujets, tailles d'effet associés aux variants causaux). Toutefois, il semble la régression *ridge* soit une méthode moins recommandée pour la confirmation d'associations PGt car elle montre une tendance à sélectionner plus de vrais comme de faux positifs. La méthode HyperLasso semble elle plus appropriée lorsque le nombre de polymorphismes à tester est très important (Hoggart et al., 2008), comme dans le cas des études GWAS. Nos conditions de simulations, se basant sur une approche gènes candidats, avec un nombre de polymorphismes plus faible, n'ont peut-être pas favorisées cette méthode par rapport aux autres.

Comparée à des études de simulations en pharmacométrie où des covariables physiopathologiques sont souvent simulées sans corrélation, une originalité de ces travaux est d'avoir simulé des variables corrélées, une caractéristique des variants génétiques. Le traitement des variants causaux, *i.e.* ceux affectant le phénotype, est aussi une particularité de nos travaux. Dans un travail de simulation, Bertrand et Balding (Bertrand and Balding, 2013) ont également simulé des polymorphismes, dont certains affectaient la pharmacocinétique d'une molécule. Ces variants causaux étaient ensuite retirés des jeux de données pour l'analyse. Dans leur analyse, tous les variants sélectionnés et corrélés avec le variant causal étaient considérés comme un signal positif. Alors que Bertrand et Balding utilisaient les corrélations entre les variables pour détecter le variant causal, nous cherchions à nous assurer que les tests d'association soient capables de détecter ce variant causal malgré la structure de corrélation. Pour cela, nous conservions les variants causaux dans les jeux de données simulés et seule la sélection de ces variants était considérée comme un signal positif. Les résultats de nos simulations montrent que les variants causaux peuvent être détectés malgré les corrélations, s'ils sont associés à des effets suffisamment forts, *i.e.* ceux expliquant la plus forte part de la variabilité de la pharmacocinétique.

Nos simulations ont également permis de montrer que le choix d'étudier la pharmacogénétique du médicament lors du développement précoce (phase I) ou avancé (phase II et phase III) affecte la puissance de l'analyse. En effet, la probabilité de détecter des effets génétiques varie en fonction du nombre de sujets et du protocole de prélèvement. Les agences du médicament comme l'EMA ou la FDA recommandent un nombre important de sujets dans les études en vue d'une analyse PGt (European Medicines Agency (EMA), 2012; Food and Drug Administration (FDA), 2013). Nos simulations confirment que le nombre de sujets est le principal paramètre influençant la puissance de détection des variants

génétiques. Mais ces simulations montrent également que la quantité d'information au niveau individuel a un effet sur cette puissance. L'intérêt des MNLEM est qu'ils permettent d'analyser les données de sujets avec un protocole éparé. Ces protocoles, s'ils ne sont pas correctement déterminés, peuvent conduire à des phénomènes de régression vers la moyenne ou *shrinkage* réduisant la puissance de détection. Nos simulations montrent qu'avec un protocole plus informatif, en augmentant sensiblement le nombre d'observations par sujet et en optimisant les temps de prélèvements, la probabilité de détecter des variants génétiques augmente.

Une nouvelle approche basée sur l'utilisation de la distribution complète des paramètres individuels dans les MNLEM a enfin été évaluée dans cette thèse. Elle permet une amélioration limitée de la puissance de détection, inférieure à celle obtenue grâce à l'optimisation des protocoles.

Les possibilités infinies qu'offrent les études de simulations obligent à fixer un cadre. Dans cette thèse nous avons limité les conditions de simulations, tant sur la simulation des profils PK que sur celles de l'effet de la génétique. Ainsi, nous avons simulé uniquement des relations additives entre polymorphismes causaux et phénotypes PK. Dans un cas réel, le fait de tester des effets uniquement additifs peut diminuer la puissance de l'étude et il est donc recommandé d'étudier aussi des relations dominantes ou récessives (Lettre et al., 2007).

Dans les différents travaux et pour chaque jeu de données simulé, les variants causaux étaient tirés aléatoirement et donc différents d'un jeu de données à l'autre. Ainsi, différents coefficients d'effet étaient simulés, car calculés notamment en fonction de la fréquence de l'allèle muté pour chaque variant causal. Cette approche est régulièrement utilisée dans les études de simulations en génétique statistique et permet aux résultats de ne pas dépendre d'une fréquence donnée (*e.g.* (Huang et al., 2014; Kichaev et al., 2014; Schifano et al., 2012)). Une alternative, plus usuelle pour des études de simulations en pharmacométrie, serait de sélectionner six variants causaux, identiques pour tous les jeux de données, et de fixer leurs coefficients d'effet associés.

Concernant les méthodes d'association, nous nous sommes également limités à quelques méthodes, cette thèse n'ayant pas pour but d'évaluer de façon exhaustive toutes les méthodes d'association en pharmacogénétique. Nous nous sommes intéressés à des méthodes basées sur une estimation par maximum de vraisemblance, dont deux méthodes

très utilisées dans la littérature (régression *ridge*, Lasso) et une méthode récente qui semblait apporter un intérêt en matière de performance de détection des variables génétiques à grande échelle (HyperLasso). Il existe dans le cadre des régressions pénalisées d'autres approches comme l'Elastic Net (Zou and Hastie, 2005) qui a montré des performances intermédiaires entre les méthodes de régression *ridge* et Lasso. En dehors des approches basées sur le maximum de vraisemblance, on trouve aussi des approches bayésiennes (O'Hara and Sillanpää, 2009) qu'il serait intéressant d'explorer. Il existe également des approches plus complexes prenant en compte les interactions gènes-environnement (Knights et al., 2013). En effet, la génétique ne permet d'expliquer qu'une part de la variabilité des processus PK. Une autre partie de cette variabilité peut être associée à des facteurs environnementaux ayant un impact directement sur le phénotype, mais également sur l'expression des gènes.

Nous ne nous sommes intéressés qu'à la pharmacocinétique des médicaments, occultant sa relation avec la pharmacodynamie. Cela est dû en partie au choix du cas réel où les données d'une puce à ADN, spécialement développée pour les études PK, ont été utilisées comme base à la simulation des polymorphismes. Les conclusions de cette thèse sur l'intérêt de la modélisation ou des protocoles combinés peuvent s'appliquer à l'étude de la pharmacodynamie. La puissance des méthodes d'association peut, elle, varier selon le type de phénotype étudié. Les résultats de cette thèse devraient être directement transposables à des phénotypes PD continus, mais demanderaient à être étendus dans le cas de phénotypes PD binaires ou catégoriels.

Concernant le modèle PK utilisé pour la simulation des phénotypes, aucune variabilité interoccasion (IOV) n'a été simulée. Cela aurait pu concerner les simulations des études de phase II où la pharmacocinétique était simulée à l'état d'équilibre. Il serait nécessaire, dans un cas réel où les sujets de l'étude sont suivis sur une longue période, de quantifier cette IOV. L'approche basée sur les MNLEM permet d'estimer ce nouveau niveau de variabilité (comme expliqué dans le paragraphe 1.1.5.1) et ainsi de la distinguer de la variabilité interindividuelle. Les paramètres associés à la variabilité interindividuelle étant, avec cette approche, estimés sans biais, des résultats similaires devraient être trouvés quant à la probabilité de détecter les variants génétiques. L'ajout d'un niveau de variabilité comme l'IOV doit également être pris en compte au moment du choix du protocole de prélèvement, afin d'avoir l'information suffisante pour l'estimation de ces nouveaux paramètres. Le

logiciel d'optimisation PFIM intègre des modèles comprenant de l'IOV pour l'optimisation des protocoles ([www.pfim.biostat.fr](http://www.pfim.biostat.fr)).

Pour de futurs travaux de simulations, il serait intéressant de simuler des effets génétiques sur plusieurs paramètres du modèle, corrélés entre eux, ce qui rajouterait un degré de difficulté pour la détection de ces polymorphismes.

Les tests d'association basés sur une régression pénalisée ne peuvent à l'heure actuelle n'être appliqués que sur les estimations individuelles du modèle. Une alternative serait d'utiliser une approche intégrant ces régressions pénalisées où les paramètres du modèle sont estimés et les covariables sont sélectionnées en une seule étape. Cette approche ne semble pour le moment pas donner de meilleures performances qu'une *stepwise procedure* alors qu'elle nécessite de plus long temps de calcul (Bertrand et al., 2015).

A l'image du programme PsN qui inclue de nombreuses fonctions pour le développement de MNLEM avec le logiciel NONMEM, il pourrait être également envisagé un programme regroupant différents modules dont un serait l'exploration des relations entre covariables et EBE *via* une régression pénalisée. Les covariables sélectionnées étant ensuite automatiquement incluses dans le modèle.

## 7.2. Recommandations pour les analyses pharmacogénétiques

Ces travaux de simulations nous permettent de fournir des recommandations afin d'augmenter la puissance des analyses PGt effectuées lors des études PK.

La NCA est, d'après la littérature, encore aujourd'hui l'approche la plus utilisée pour estimer les paramètres PK associés aux polymorphismes génétiques dans les analyses PGt, certainement pour sa facilité de mise en œuvre. Nos travaux montrent que la modélisation et plus particulièrement les MNLEM doivent pourtant être préférés, car ils offrent de meilleures performances pour les analyses PGt.

Les différents tests d'association étudiés montrent des performances similaires. La méthode *stepwise procedure* qui est usuellement utilisée en pharmacométrie peut donc être recommandée pour les analyses PGt. Cette méthode peut tout de même nécessiter un nombre important d'itérations et donc un temps de calcul élevé dans le cas où beaucoup de relations sont explorées. Les approches basées sur une régression pénalisée ont dans ce cas

l'avantage de fournir un modèle final plus rapidement, en explorant tous les variants simultanément.

Nos simulations montrent l'intérêt de combiner les données de sujets avec des protocoles de prélèvement épars avec des données de protocoles plus riches chez un sous-groupe de sujets, afin d'augmenter la quantité d'information et donc la puissance de l'étude. L'optimisation de ces protocoles épars *via* des algorithmes et des logiciels spécifiques, *e.g.* le logiciel PFIM ([www.pfim.biostat.fr](http://www.pfim.biostat.fr)), apporte de plus un vrai bénéfice sur la puissance de détection. Nous recommandons donc sur la base de ces résultats d'étudier la pharmacogénétique d'une molécule lors de son développement sur des données combinant les études de phase I et de phase II, et optimisées prospectivement.

Au niveau du développement de nouveaux médicaments, l'étude de la pharmacogénétique pourrait apporter de vrais bénéfices alors que les coûts de développement sont en augmentation (Kaitin, 2010) et qu'à l'inverse les taux de succès ont tendance à diminuer (Pammolli et al., 2011). Des problèmes liés à l'efficacité sont majoritairement responsables de l'arrêt du développement des nouveaux traitements (Kola and Landis, 2004). À l'heure actuelle, il existe peu d'exemple montrant l'apport des analyses PGt dans le développement de nouveaux médicaments, comparé aux nombreuses associations mises en évidence et ayant abouti à des recommandations d'usage en pratique clinique pour des médicaments déjà sur le marché. L'utilisation de marqueurs PGt permettrait notamment de relancer le développement de certaines molécules précédemment arrêtées par manque d'efficacité ou à cause d'effets indésirables trop importants (Shah, 2006). Par exemple, ce fut le cas du lumiracoxib, traitement de la douleur aiguë et de l'ostéoarthrite, dont le développement fut arrêté en 2005 en raison de toxicités hépatiques. Des analyses pharmacogénétiques rétrospectives ont montré que des variants du gène HLA-DQ étaient associés à une élévation des enzymes hépatiques et qu'un allèle (HLA-DQA1\*0102) était présent chez 100% des patients présentant les niveaux d'enzymes les plus élevés (Singer et al., 2010). Le développement de cette molécule a donc été relancé avec une restriction à une population sélectionnée sur la génétique.

Les associations PGt identifiées en phase précoce du développement de nouveaux médicaments (phase I et phase II) permettraient de mieux concevoir les études de phase III. Par exemple en adaptant la dose en fonction du génotype ou en excluant les patients non répondeurs, l'efficacité estimée de ces nouveaux médicaments pourrait augmenter et le

taux d'effets indésirables diminuer. L'intérêt de ces protocoles basés sur la génétique est de rendre le développement clinique de nouveaux médicaments plus efficace, en matière de coûts et de temps de développement (Cook et al., 2009), et d'en augmenter le taux de succès.

La pharmacogénétique change le paradigme de la recherche clinique dans le domaine du médicament : celui d'un traitement unique capable de traiter l'ensemble de la population. Aujourd'hui avec le développement de la pharmacogénétique et de la médecine personnalisée on cherche le médicament capable de traiter des sous-groupes de patients, voire à la limite le patient unique.

## BIBLIOGRAPHIE

1. Aarons, L., 2005. Physiologically based pharmacokinetic modelling: a sound mechanistic basis is needed. *Br. J. Clin. Pharmacol.* 60, 581–583.
2. Aithal, G.P., Day, C.P., Kesteven, P.J., Daly, A.K., 1999. Association of polymorphisms in the cytochrome P450 CYP2C9 with warfarin dose requirement and risk of bleeding complications. *Lancet* 353, 717–719.
3. Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19, 716–723.
4. Allorge, D., Lorient, M.-A., 2004. [Pharmacogenetics or the promise of a personalized medicine: variability in drug metabolism and transport]. *Ann. Biol. Clin. (Paris)* 62, 499–511.
5. Balding, D.J., 2006. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* 7, 781–791.
6. Bazzoli, C., Retout, S., Mentré, F., 2010. Design evaluation and optimisation in multiple response nonlinear mixed effect models: PFIM 3.0. *Comput. Methods Programs Biomed.* 98, 55–65.
7. Beal, S., Sheiner, L.B., Boeckmann, A., Bauer, R.J., 2009. NONMEM User's guides. Icon development Solutions, Ellicott City, USA.
8. Becquemont, L., Alfirevic, A., Amstutz, U., Brauch, H., Jacqz-Aigrain, E., Laurent-Puig, P., Molina, M.A., Niemi, M., Schwab, M., Somogyi, A.A., et al, 2011. Practical recommendations for pharmacogenomics-based prescription: 2010 ESF-UB Conference on Pharmacogenetics and Pharmacogenomics. *Pharmacogenomics* 12, 113–124.
9. Bellman, R., Åström, K.J., 1970. On structural identifiability. *Math. Biosci.* 7, 329–339.
10. Bender, R., Lange, S., 2001. Adjusting for multiple testing--when and how? *J. Clin. Epidemiol.* 54, 343–349.
11. Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.
12. Bergstrand, M., Hooker, A.C., Wallin, J.E., Karlsson, M.O., 2011. Prediction-Corrected Visual Predictive Checks for Diagnosing Nonlinear Mixed-Effects Models. *AAPS J.* 13, 143–151.
13. Bertrand, J., 2009. Pharmacogénétique en pharmacocinétique de population : tests et sélection de modèles. Paris Diderot (thèse d'université).
14. Bertrand, J., Balding, D.J., 2013. Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models. *Pharmacogenet. Genomics* 23, 167–174.
15. Bertrand, J., Comets, E., Laffont, C.M., Chenel, M., Mentré, F., 2009. Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm. *J. Pharmacokinet. Pharmacodyn.* 36, 317–339.



16. Bertrand, J., De Iorio, M., Balding, D.J., 2015. Integrating dynamic mixed-effect modelling and penalized regression to explore genetic association with pharmacokinetics. *Pharmacogenet. Genomics* 25, 231–238.
17. Bertrand, J., Verstuyft, C., Chou, M., Borand, L., Chea, P., Nay, K.H., Blanc, F.-X., Mentré, F., Taburet, A.-M., CAMELIA (ANRS 1295-CIPRA KH001) Study Group, 2014. Dependence of efavirenz- and rifampicin-isoniazid-based antituberculosis treatment drug-drug interaction on CYP2B6 and NAT2 genetic polymorphisms: ANRS 12154 study in Cambodia. *J. Infect. Dis.* 209, 399–408.
18. Bland, J.M., Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. *BMJ* 310, 170.
19. Botton, M.R., Bandinelli, E., Rohde, L.E.P., Amon, L.C., Hutz, M.H., 2011. Influence of genetic, biological and pharmacological factors on warfarin dose in a Southern Brazilian population of European ancestry. *Br. J. Clin. Pharmacol.* 72, 442–450.
20. Brendel, K., Comets, E., Laffont, C., Laveille, C., Mentre, F., 2006. Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. *Pharm. Res.* 23, 2036–2049.
21. Brendel, K., Dartois, C., Comets, E., Lemenuel-Diot, A., Laveille, C., Tranchand, B., Girard, P., Laffont, C.M., Mentré, F., 2007. Are population pharmacokinetic and/or pharmacodynamic models adequately evaluated? A survey of the literature from 2002 to 2004. *Clin. Pharmacokinet.* 46, 221–234.
22. Burton, M.E., Shaw, L.M., Schentag, J.J., Evans, W.E. (Eds.), 2005. *Applied Pharmacokinetics and Pharmacodynamics: Principles of Therapeutic Drug Monitoring*, Fourth edition. ed. Lippincott William & Wilkins, Baltimore.
23. Bush, W.S., Moore, J.H., 2012. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* 8, e1002822.
24. Cappellini, M., Fiorelli, G., 2008. Glucose-6-phosphate dehydrogenase deficiency. *The Lancet* 371, 64–74.
25. Chatelut, E., Canal, P., Brunner, V., Chevreau, C., Pujol, A., Boneu, A., Roché, H., Houin, G., Bugat, R., 1995. Prediction of carboplatin clearance from standard morphological and biological patient characteristics. *J. Natl. Cancer Inst.* 87, 573–580.
26. Chevance, F.F.V., Le Guyon, S., Hughes, K.T., 2014. The Effects of Codon Context on In Vivo Translation Speed. *PLoS Genet* 10, e1004392.
27. Combes, F., Retout, S., Frey, N., Mentré, F., 2014. Powers of the Likelihood Ratio Test and the Correlation Test Using Empirical Bayes Estimates for Various Shrinkages in Population Pharmacokinetics. *CPT Pharmacometrics Syst. Pharmacol.* 3, 1–9.
28. Comets, E., Lavenu, A., Lavielle, M., 2011. saemix: Stochastic Approximation Expectation Maximization (SAEM) algorithm. R package version 1.0.
29. Comets, E., Verstuyft, C., Lavielle, M., Jaillon, P., Becquemont, L., Mentré, F., 2007. Modelling the influence of MDR1 polymorphism on digoxin pharmacokinetic parameters. *Eur. J. Clin. Pharmacol.* 63, 437–449.

30. Commenges, D., Jacqmin-Gadda, H., Proust, C., Guedj, J., 2006. A Newton-Like Algorithm for Likelihood Maximization: The Robust-Variance Scoring Algorithm. *Arxiv Math0610402*.
31. Cook, J., Hunter, G., Vernon, J.A., 2009. The future costs, risks and rewards of drug development: the economics of pharmacogenomics. *Pharmacoeconomics* 27, 355–363.
32. Cule, E., De Iorio, M., 2013. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet. Epidemiol.* 37, 704–714.
33. Cule, E., Vineis, P., De Iorio, M., 2011. Significance testing in ridge regression for genetic data. *BMC Bioinformatics* 12, 372.
34. de Keyser, C.E., Peters, B.J.M., Becker, M.L., Visser, L.E., Uitterlinden, A.G., Klungel, O.H., Verstuyft, C., Hofman, A., Maitland-van der Zee, A.-H., Stricker, B.H., 2014. The SLCO1B1 c.521T>C polymorphism is associated with dose decrease or switching during statin therapy in the Rotterdam Study. *Pharmacogenet. Genomics* 24, 43–51.
35. de Zwart, L.L., Haenen, H.E.M.G., Versantvoort, C.H.M., Wolterink, G., van Engelen, J.G.M., Sips, A.J. a. M., 2004. Role of biokinetics in risk assessment of drugs and chemicals in children. *Regul. Toxicol. Pharmacol. RTP* 39, 282–309.
36. Dichgans, M., Markus, H.S., 2005. Genetic association studies in stroke: methodological issues and proposed standard criteria. *Stroke J. Cereb. Circ.* 36, 2027–2031.
37. Dost, F.H., 1953. *Der Blütspiegel-Kinetic der Konzentrationsabläufe in der Krieslauffüssigkeit*. G. Thieme, Leipzig.
38. Efron, B., Morris, C., 1977. Stein's Paradox in Statistics. *Sci. Am.* 236, 119–127.
39. Efron, B., Morris, C., 1971. Limiting the Risk of Bayes and Empirical Bayes Estimators--Part I: The Bayes Case. *J. Am. Stat. Assoc.* 66, 807–815.
40. Eichelbaum, M., Ingelman-Sundberg, M., Evans, W.E., 2006. Pharmacogenomics and Individualized Drug Therapy. *Annu. Rev. Med.* 57, 119–137.
41. European Medicines Agency (EMA), 2012. Guideline on the use of pharmacogenetic methodologies in the pharmacokinetic evaluation of medicinal products (No. EMA/CHMP/37646/2009).
42. European Medicines Agency (EMA), 2006. Guideline on reporting the results of population pharmacokinetic analyses (No. EMA/CHMP/EWP/185990/2006).
43. Evans, W.E., Relling, M.V., 1999. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 286, 487–491.
44. Food and Drug Administration (FDA), 2013. Clinical Pharmacogenomics: Premarket Evaluation in Early-Phases Clinical Studies and Recommendations for Labeling.
45. Food and Drug Administration (FDA), 2008. E15 Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories.
46. Food and Drug Administration (FDA), 2003. Exposure-Response Relationships — Study Design, Data Analysis, and Regulatory Applications.

47. Frueh, F.W., Amur, S., Mummaneni, P., Epstein, R.S., Aubert, R.E., DeLuca, T.M., Verbrugge, R.R., Burckart, G.J., Lesko, L.J., 2008. Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. *Pharmacotherapy* 28, 992–998.
48. Gabrielson, J., Weiner, D., 2007. *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications*, Fourth Edition. ed. Swedish Pharmaceutical Press.
49. Galetin, A., Gertz, M., Houston, J.B., 2010. Contribution of intestinal cytochrome p450-mediated metabolism to drug-drug inhibition and induction interactions. *Drug Metab. Pharmacokinet.* 25, 28–47.
50. Giacomini, K.M., Huang, S.-M., 2013. Transporters in drug development and clinical pharmacology. *Clin. Pharmacol. Ther.* 94, 3–9.
51. Gibaldi, M., Levy, G., 1976. Pharmacokinetics in clinical practice: I. concepts. *JAMA* 235, 1864–1867.
52. Godfrey, K.R., Chapman, M.J., Vajda, S., 1994. Identifiability and indistinguishability of nonlinear pharmacokinetic models. *J. Pharmacokinet. Biopharm.* 22, 229–251.
53. Gradshteyn, I., Ryzik, I., 1980. *Tables of Integrals, Series and Products: Corrected and Enlarged Edition.*, New York: Academic Press. ed.
54. Guedj, J., Thiébaud, R., Commenges, D., 2007. Maximum Likelihood Estimation in Dynamical Models of HIV. *Biometrics* 63, 1198–1206.
55. Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 55.
56. Hoggart, C.J., Whittaker, J.C., De Iorio, M., Balding, D.J., 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* 4, e1000130.
57. Holt, D.W., Tucker, G.T., Jackson, P.R., Storey, G.C.A., 1983. Amiodarone pharmacokinetics. *Am. Heart J.*, Amiodarone: Basic Concepts and Clinical Applications 106, 840–847.
58. Hong, E.P., Park, J.W., 2012. Sample size and statistical power calculation in genetic association studies. *Genomics Inform.* 10, 117–122.
59. Houin, G., 1998. Principes pharmacocinétiques de l'adaptation de la posologie: définitions et sources de variabilité. *Rev. Fr. Lab.* 1998, 25–31.
60. Huang, Y.-T., VanderWeele, T.J., Lin, X., 2014. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Ann. Appl. Stat.* 8, 352–376.
61. Hubbard, T.J.P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al, 2007. Ensembl 2007. *Nucleic Acids Res.* 35, D610–617.
62. Huque, M.F., Sankoh, A.J., 1997. A reviewer's perspective on multiple endpoint issues in clinical trials. *J. Biopharm. Stat.* 7, 545–564.
63. Ingelman-Sundberg, M., 2008. Pharmacogenomic Biomarkers for Prediction of Severe Adverse Drug Reactions. *N. Engl. J. Med.* 358, 637–639.
64. International HapMap Consortium, 2003. The International HapMap Project. *Nature* 426, 789–796.

65. International Warfarin Pharmacogenetics Consortium, Klein, T.E., Altman, R.B., Eriksson, N., Gage, B.F., Kimmel, S.E., Lee, M.-T.M., Limdi, N.A., Page, D., Roden, D.M., et al, 2009. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* 360, 753–764.
66. James, W., Stein, C., 1961. Estimation with Quadratic Loss. Presented at the Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, The Regents of the University of California.
67. Johnson, T.N., Rostami-Hodjegan, A., Tucker, G.T., 2006. Prediction of the clearance of eleven drugs and associated variability in neonates, infants and children. *Clin. Pharmacokinet.* 45, 931–956.
68. Joy, T.R., Hegele, R.A., 2009. Narrative review: statin-related myopathy. *Ann. Intern. Med.* 150, 858–868.
69. Kaitin, K.I., 2010. Deconstructing the drug development process: the new face of innovation. *Clin. Pharmacol. Ther.* 87, 356–361.
70. Kalliokoski, A., Niemi, M., 2009. Impact of OATP transporters on pharmacokinetics. *Br. J. Pharmacol.* 158, 693–705.
71. Karlsson, M.O., Sheiner, L.B., 1993. The importance of modeling interoccasion variability in population pharmacokinetic analyses. *J. Pharmacokinet. Biopharm.* 21, 735–750.
72. Kauf, T.L., Farkouh, R.A., Earnshaw, S.R., Watson, M.E., Maroudas, P., Chambers, M.G., 2010. Economic efficiency of genetic screening to inform the use of abacavir sulfate in the treatment of HIV. *PharmacoEconomics* 28, 1025–1039.
73. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., Pasaniuc, B., 2014. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genet.* 10.
74. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al, 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385–389.
75. Knights, J., Chanda, P., Sato, Y., Kaniwa, N., Saito, Y., Ueno, H., Zhang, A., Ramanathan, M., 2013. Vertical Integration of Pharmacogenetics in Population PK/PD Modeling: A Novel Information Theoretic Method. *CPT Pharmacometrics Syst. Pharmacol.* 2, e25.
76. Kola, I., Landis, J., 2004. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715.
77. Kuhn, E., Lavielle, M., 2004. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat.* 8, 115–131.
78. Lai, T.L., Shih, M.-C., 2003. Nonparametric Estimation in Nonlinear Mixed Effects Models. *Biometrika* 90, 1–13.
79. Lamba, J., Lamba, V., Strom, S., Venkataramanan, R., Schuetz, E., 2008. Novel single nucleotide polymorphisms in the promoter and intron 1 of human pregnane X receptor/NR1I2 and their association with CYP3A4 expression. *Drug Metab. Dispos. Biol. Fate Chem.* 36, 169–181.
80. Lander, E.S., 1999. Array of hope. *Nat. Genet.* 21, 3–4.

81. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al, 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
82. Lavielle M, Mesa H, Chatel K, 2010. The MONOLIX software.
83. Lay, M.J., Wittwer, C.T., 1997. Real-time fluorescence genotyping of factor V Leiden during rapid-cycle PCR. *Clin. Chem.* 43, 2262–2267.
84. Lehr, T., Schaefer, H.-G., Staab, A., 2010. Integration of high-throughput genotyping data into pharmacometric analyses using nonlinear mixed effects modeling. *Pharmacogenet. Genomics* 20, 442–450.
85. Lettre, G., Lange, C., Hirschhorn, J.N., 2007. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.* 31, 358–362.
86. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al, 2007. The diploid genome sequence of an individual human. *PLoS Biol.* 5, e254.
87. Lewontin, R.C., 1988. On measures of gametic disequilibrium. *Genetics* 120, 849–852.
88. Lindbom, L., Pihlgren, P., Jonsson, E.N., Jonsson, N., 2005. PsN-Toolkit--a collection of computer intensive statistical methods for non-linear mixed effect modeling using NONMEM. *Comput. Methods Programs Biomed.* 79, 241–257.
89. Lindbom, L., Ribbing, J., Jonsson, E.N., 2004. Perl-speaks-NONMEM (PsN)--a Perl module for NONMEM related programming. *Comput. Methods Programs Biomed.* 75, 85–94.
90. Lindstrom, M.L., Bates, D.M., 1990. Nonlinear mixed effects models for repeated measures data. *Biometrics* 46, 673–687.
91. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., et al, 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675–1680.
92. Lopez-Lopez, E., Martin-Guerrero, I., Ballesteros, J., Piñan, M.A., Garcia-Miguel, P., Navajas, A., Garcia-Orad, A., 2011. Polymorphisms of the SLCO1B1 gene predict methotrexate-related toxicity in childhood acute lymphoblastic leukemia. *Pediatr. Blood Cancer* 57, 612–619.
93. Maliepaard, M., Nofziger, C., Papaluca, M., Zineh, I., Uyama, Y., Prasad, K., Grimstein, C., Pacanowski, M., Ehmann, F., Dossena, S., et al, 2013. Pharmacogenetics in the evaluation of new drugs: a multiregional regulatory perspective. *Nat. Rev. Drug Discov.* 12, 103–115.
94. Mallal, S., Phillips, E., Carosi, G., Molina, J.-M., Workman, C., Tomazic, J., Jägel-Guedes, E., Rugina, S., Kozyrev, O., Cid, J.F., et al, 2008. HLA-B\*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* 358, 568–579.
95. Mallet, A., 1986. A maximum likelihood estimation method for random coefficient regression models. *Biometrika* 73, 645–656.

96. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al, 2009. Finding the missing heritability of complex diseases. *Nature* 461, 747–753.
97. Marzolini, C., Paus, E., Buclin, T., Kim, R.B., 2004. Polymorphisms in human MDR1 (P-glycoprotein): recent advances and clinical relevance. *Clin. Pharmacol. Ther.* 75, 13–33.
98. Mathijssen, R.H.J., Sparreboom, A., Verweij, J., 2014. Determining the optimal dose in the development of anticancer agents. *Nat. Rev. Clin. Oncol.* 11, 272–281.
99. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., Hirschhorn, J.N., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369.
100. Metzker, M.L., 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46.
101. Michaelis, L., Menten, M.L., 1913. Die Kinetik der Invertinwirkung. *Biochem. Z.* 49, 333–369.
102. Michaelis, L., Menten, M.L., Johnson, K.A., Goody, R.S., 2011. The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry (Mosc.)* 50, 8264–8269.
103. Motulsky, A.G., 1957. Drug reactions, enzymes, and biochemical genetics. *J. Am. Med. Assoc.* 165, 835–837.
104. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., Erlich, H., 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* 51 Pt 1, 263–273.
105. Nelson, E., 1961. Kinetics of drug absorption, distribution, metabolism, and excretion. *J. Pharm. Sci.* 50, 181–192.
106. Nyholt, D.R., 2004. A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other. *Am. J. Hum. Genet.* 74, 765–769.
107. O’Hara, R.B., Sillanpää, M.J., 2009. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.* 4, 85–117.
108. Pammolli, F., Magazzini, L., Riccaboni, M., 2011. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* 10, 428–438.
109. Pillai, G.C., Mentré, F., Steimer, J.-L., 2005. Non-linear mixed effects modeling - from methodology and software development to driving implementation in drug development science. *J. Pharmacokinet. Pharmacodyn.* 32, 161–183.
110. Pinheiro, J., Bates, D., 2000. *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media.
111. Prague, M., Commenges, D., Guedj, J., Drylewicz, J., Thiébaud, R., 2013. NIMROD: a program for inference via a normal approximation of the posterior in models with random effects based on ordinary differential equations. *Comput. Methods Programs Biomed.* 111, 447–458.

112. Pritchard, J.K., Cox, N.J., 2002. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum. Mol. Genet.* 11, 2417–2423.
113. Ramsey, L.B., Bruun, G.H., Yang, W., Treviño, L.R., Vattathil, S., Scheet, P., Cheng, C., Rosner, G.L., Giacomini, K.M., Fan, Y., et al, 2012. Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition. *Genome Res.* 22, 1–8.
114. R Development Core Team, 2012. A language and environment for statistical computing. R Foundation for Statistical Computing.
115. Ribbing, J., Nyberg, J., Caster, O., Jonsson, E.N., 2007. The lasso--a novel method for predictive covariate model building in nonlinear mixed effects models. *J. Pharmacokinet. Pharmacodyn.* 34, 485–517.
116. Rieder, M.J., Reiner, A.P., Gage, B.F., Nickerson, D.A., Eby, C.S., McLeod, H.L., Blough, D.K., Thummel, K.E., Veenstra, D.L., Rettie, A.E., 2005. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N. Engl. J. Med.* 352, 2285–2293.
117. Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science* 273, 1516–1517.
118. Robert, C.P., 1994. *The Bayesian Choice*, Springer Texts in Statistics. Springer-Verlag, New York, NY.
119. Rosell, R., Moran, T., Queralt, C., Porta, R., Cardenal, F., Camps, C., Majem, M., Lopez-Vivanco, G., Isla, D., Provencio, M., et al, 2009. Screening for epidermal growth factor receptor mutations in lung cancer. *N. Engl. J. Med.* 361, 958–967.
120. Rowland, M., Noe, C.R., Smith, D.A., Tucker, G.T., Crommelin, D.J.A., Peck, C.C., Rocci, M.L., Besançon, L., Shah, V.P., 2012. Impact of the pharmaceutical sciences on health care: a reflection over the past 50 years. *J. Pharm. Sci.* 101, 4075–4099.
121. Rowland, M., Tozer, T.N., 2011. *Clinical pharmacokinetics and pharmacodynamics: concepts and applications*. Wolters Kluwer Health/Lippincott William & Wilkins, Philadelphia.
122. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., Smith, M., 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687–695.
123. SAS Institute Inc., 2002. *SAS 9.1.3 Help and Documentation*. Cary, NC: SAS Institute Inc.
124. Sauna, Z.E., Kimchi-Sarfaty, C., 2011. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* 12, 683–691.
125. Savic, R.M., Karlsson, M.O., 2009. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *AAPS J.* 11, 558–569.
126. Schifano, E.D., Epstein, M.P., Bielak, L.F., Jhun, M.A., Kardia, S.L.R., Peyser, P.A., Lin, X., 2012. SNP Set Association Analysis for Familial Data. *Genet. Epidemiol.* 36, 797–810.
127. Schroth, W., Goetz, M.P., Hamann, U., Fasching, P.A., Schmidt, M., Winter, S., Fritz, P., Simon, W., Suman, V.J., Ames, M.M., et al, 2009. Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. *JAMA J. Am. Med. Assoc.* 302, 1429–1436.

128. Schwarz, G., 1978. Estimating the Dimension of a Model. *Ann. Stat.* 6, 461–464.
129. Sconce, E.A., Khan, T.I., Wynne, H.A., Avery, P., Monkhouse, L., King, B.P., Wood, P., Kesteven, P., Daly, A.K., Kamali, F., 2005. The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. *Blood* 106, 2329–2333.
130. SEARCH Collaborative Group, Link, E., Parish, S., Armitage, J., Bowman, L., Heath, S., Matsuda, F., Gut, I., Lathrop, M., Collins, R., 2008. SLCO1B1 variants and statin-induced myopathy--a genomewide study. *N. Engl. J. Med.* 359, 789–799.
131. Shah, R.R., 2006. Can pharmacogenetics help rescue drugs withdrawn from the market? *Pharmacogenomics* 7, 889–908.
132. Sheiner, L.B., 1984. The population approach to pharmacokinetic data analysis: rationale and standard data analysis methods. *Drug Metab. Rev.* 15, 153–171.
133. Sheiner, L.B., Beal, S.L., 1980. Evaluation of methods for estimating population pharmacokinetics parameters. I. Michaelis-Menten model: routine clinical pharmacokinetic data. *J. Pharmacokinet. Biopharm.* 8, 553–571.
134. Sheiner, L.B., Rosenberg, B., Melmon, K.L., 1972. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput. Biomed. Res. Int. J.* 5, 411–459.
135. Shou, W., Wang, D., Zhang, K., Wang, B., Wang, Z., Shi, J., Huang, W., 2012. Gene-Wide Characterization of Common Quantitative Trait Loci for ABCB1 mRNA Expression in Normal Liver Tissues in the Chinese Population. *PLoS ONE* 7, e46295.
136. Šidák, Z., 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *J. Am. Stat. Assoc.* 62, 626–633.
137. Sim, S.C., Ingelman-Sundberg, M., 2011. Pharmacogenomic biomarkers: new tools in current and future drug therapy. *Trends Pharmacol. Sci.* 32, 72–81.
138. Singer, J.B., Lewitzky, S., Leroy, E., Yang, F., Zhao, X., Klickstein, L., Wright, T.M., Meyer, J., Paulding, C.A., 2010. A genome-wide study identifies HLA alleles associated with lumiracoxib-related liver injury. *Nat. Genet.* 42, 711–714.
139. Steimer, J.-L., 1992. Population models and methods, with emphasis on pharmacokinetics, in: Rowland M., Aarons L. (eds). *New Strategies in Drug Development and Clinical Evaluation: The Population Approach*. Commission of the European Communities, Luxembourg.
140. Steimer, J.-L., Vozeh, S., Racine-Poon, A., Holford, N., O'Neill, R., 1994. The Population Approach: Rationale, Methods, and Applications in Clinical Pharmacology and Drug Development, in: D.Sc, P.G.W., Balant, L.P. (Eds.), *Pharmacokinetics of Drugs, Handbook of Experimental Pharmacology*. Springer Berlin Heidelberg, pp. 405–451.
141. Stein, C., 1956. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. Presented at the Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, The Regents of the University of California.
142. Suarez-Kurtz, G., 2011. Population diversity and the performance of warfarin dosing algorithms. *Br. J. Clin. Pharmacol.* 72, 451–453.



143. Suarez-Kurtz, G., 2010. Pharmacogenetics in the brazilian population. *Front. Pharmacol.* 1, 118.
144. Teorell, T., 1937. Kinetics of distribution of substances administered to the body. *Arch Intern Pharmacodyn* 57, 205–40.
145. The 1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65.
146. The International HapMap Consortium, 2005. A haplotype map of the human genome. *Nature* 437, 1299–1320.
147. Tibshirani, R., 1994. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288.
148. Tozer, T.N., Rowland, M., 2006. Introduction to Pharmacokinetics and Pharmacodynamics: The Quantitative Basis of Drug Therapy. Lippincott Williams & Wilkins.
149. Tregouet, D.-A., Tired, L., 2004. Cox proportional hazards survival regression in haplotype-based association analysis using the Stochastic-EM algorithm. *Eur. J. Hum. Genet. EJHG* 12, 971–974.
150. Treviño, L.R., Shimasaki, N., Yang, W., Panetta, J.C., Cheng, C., Pei, D., Chan, D., Sparreboom, A., Giacomini, K.M., Pui, C.-H., et al, 2009. Germline Genetic Variation in an Organic Anion Transporter Polypeptide Associated With Methotrexate Pharmacokinetics and Clinical Effects. *J. Clin. Oncol.* 27, 5972–5978.
151. Verbeke, G., Molenberghs, G., 2009. Linear Mixed Models for Longitudinal Data. Springer-Verlag, New York, NY.
152. Verstuyft, C., Strabach, S., El-Morabet, H., Kerb, R., Brinkmann, U., Dubert, L., Jaillon, P., Funck-Brentano, C., Trugnan, G., Becquemont, L., 2003. Dipyridamole enhances digoxin bioavailability via P-glycoprotein inhibition. *Clin. Pharmacol. Ther.* 73, 51–60.
153. Verzelen, N., 2012. Minimax risks for sparse regressions: Ultra-high dimensional phenomena. *Electron. J. Stat.* 6, 38–90.
154. Vignal, C.M., Bansal, A.T., Balding, D.J., 2011. Using penalised logistic regression to fine map HLA variants for rheumatoid arthritis. *Ann. Hum. Genet.* 75, 655–664.
155. Vogel, F., 1959. Moderne Probleme der Humangenetik, in: Heilmeyer, L., Schoen, R., Rudder, B. de (Eds.), *Ergebnisse der Inneren Medizin und Kinderheilkunde, Ergebnisse der Inneren Medizin und Kinderheilkunde*. Springer Berlin Heidelberg, pp. 52–125.
156. Wagner, J.G., 1981. History of pharmacokinetics. *Pharmacol. Ther.* 12, 537–562.
157. Weinshilboum, R., 2003. Inheritance and Drug Response. *N. Engl. J. Med.* 348, 529–537.
158. Wendling, T., Dumitras, S., Ogungbenro, K., Aarons, L., 2015. Application of a Bayesian approach to physiological modelling of mavoglurant population pharmacokinetics. *J. Pharmacokinet. Pharmacodyn.* 42, 639–657.
159. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., et al, 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 35, D5–12.

160. Widmark, E.M.P., 1932. Die theoretischen grundlagen und die praktische verwendbarkeit der gerichtlich-medizinischen alkoholbestimmung. Urban & Schwarzenberg, Berlin, Wien.
161. Wienkers, L.C., Heath, T.G., 2005. Predicting in vivo drug interactions from in vitro drug discovery data. *Nat. Rev. Drug Discov.* 4, 825–833.
162. Wilkinson, G.R., 2005. Drug Metabolism and Variability among Patients in Drug Response. *N. Engl. J. Med.* 352, 2211–2221.
163. Wolfinger, R., 1993. Laplace's approximation for nonlinear mixed models. *Biometrika* 80, 791–795.
164. Yoshida, T., Zhang, G., Haura, E.B., 2010. Targeting epidermal growth factor receptor: central signaling kinase in lung cancer. *Biochem. Pharmacol.* 80, 613–623.
165. Zondervan, K.T., Cardon, L.R., 2007. Designing candidate gene and genome-wide case-control association studies. *Nat. Protoc.* 2, 2492–2501.
166. Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301–320.

## Résumé

La pharmacogénétique (PGt) étudie la part de la variabilité interindividuelle dans la réponse aux traitements expliquée par des variations génétiques. La PGt relie les génotypes de substitutions nucléotidiques à la variabilité pharmacocinétique (PK) d'un médicament. Dans l'objectif d'individualiser le traitement, des données génétiques sont collectées dans de nombreux essais cliniques. Il n'existe pas de consensus sur la méthodologie pour étudier l'effet de la génétique sur la PK, notamment lors du développement clinique de médicaments. Nous évaluons et comparons les méthodes d'analyse PGt dans les études PK en phases précoces du développement, afin de proposer des approches améliorant la détection des effets génétiques.

Dans une première série de simulations basée sur un exemple réel, nous comparons différentes méthodes utilisées en PGt afin de détecter l'effet simulé de plusieurs variants génétiques : les méthodes pour estimer le phénotype PK (analyse non compartimentale et modèles non linéaires à effets mixtes (MNLEM)) ; les méthodes d'association (approche itérative et trois régressions pénalisées : régression *ridge*, Lasso et HyperLasso). Dans une seconde étude de simulation nous proposons des protocoles d'étude pour augmenter la puissance de détection des variants génétiques. Enfin nous évaluons par simulations dans une troisième étude une approche pour corriger la régression vers la moyenne des phénotypes PK estimés par MNLEM qui entraîne une diminution de la puissance de détection des variants génétiques.

À travers ces différentes études de simulations, nous proposons des recommandations pour les analyses PGt lors d'étude PK dans le développement du médicament.

## Abstract

Pharmacogenetics (PGt) studies the proportion of interindividual variability in drug response explained by genetic variations. Pharmacogenetics relates especially the genotypes of single nucleotide polymorphisms to pharmacokinetic (PK) variability of a drug. In the hopes to individualise treatments, genetic data is collected in many clinical trials. There is no consensus on methodology to study the effect of genetics on PK, especially during drug development. We investigate and compare methods for PGt analyses in PK early phase studies, to propose approaches enhancing the detection of genetic effects.

In a first simulation based on a motivating example, we compare different methods used in PGt to detect the simulated effect of several genetic variants: methods to estimate the PK phenotype (noncompartmental analysis and nonlinear mixed effects models (NLMEM)); association methods (stepwise procedure and three penalised regressions: ridge regression, Lasso and HyperLasso). In a second simulation study we propose practical study designs to improve detection power of genetic variants during drug development. In a third study we assess through simulations an approach to correct the shrinkage in PK phenotype estimated through NLMEM which results in a reduced power to detect genetic variants.

Through these different simulation studies, we propose recommendations for PGt analyses in PK studies during drug development.

## Mots-clés

Pharmacogénétique ; Pharmacocinétique ; Analyses de population ; Modèles non linéaires à effets mixtes ; Régressions pénalisées ; Études cliniques ; Développement de médicaments ; Simulations